Augmenting Dementia Cognitive Assessment with Instruction-less Eye-tracking Tests

Kyriaki Mengoudi, Daniele Ravi, Keir X X Yong, Silvia Primativo, Ivanna M Pavisic, Emilie Brotherhood, Kirsty Lu, Jonathan M Schott, Sebastian J Crutch, Daniel C Alexander

***Abstract*—Eye-tracking technology is an innovative tool that holds promise for enhancing dementia screening. In this work, we introduce a novel way of extracting salient features directly from the raw eye-tracking data of a mixed sample of dementia patients during a novel instruction-less cognitive test. Our approach is based on self-supervised representation learning where, by training initially a deep neural network to solve a pretext task using well-defined available labels (e.g. recognising distinct cognitive activities in healthy individuals), the network encodes high-level semantic information which is useful for solving other problems of interest (e.g. dementia classification). Inspired by previous work in explainable AI, we use the Layer-wise Relevance Propagation (LRP) technique to describe our network's decisions in differentiating between the distinct cognitive activities. The extent to which eye-tracking features of dementia patients deviate from healthy behaviour is then explored, followed by a comparison between self-supervised and handcrafted representations on discriminating between participants with and without dementia. Our findings not only reveal novel self-supervised learning features that are more sensitive than handcrafted features in detecting performance differences between participants with and without dementia across a variety of tasks, but also validate that instruction-less eye-tracking tests can detect oculomotor biomarkers of dementia-related cognitive dysfunction. This work highlights the contribution of self-supervised representation learning techniques in biomedical applications where the small number of patients, the non-homogenous presentations of the disease and the complexity of the setting can be a challenge using state-of-the-art feature extraction methods.**

***Index Terms*—eye-tracking, dementia, cognition, deep-learning, representation learning**

I. Introduction

DEMENTIA represents a major, global healthcare challenge. This umbrella term covers a number of neurodegenerative syndromes featuring gradual disturbance of various cognitive functions that are severe enough to interfere with tasks of daily life. Alzheimer's disease (AD), the most common cause of dementia, is most commonly characterised by gradual episodic memory impairment whereas atypical forms may primarily affect vision (posterior cortical atrophy; PCA), language (logopenic variant primary progressive aphasia; lvPPA) or behaviour and executive functions. Other diseases can lead to language-led dementias including semantic variant (svPPA) and progressive non fluent variants of primary progressive aphasia (nfvPPA), while subtypes led by behavioural change include the behavioural variant of frontotemporal dementia (bvFTD).

Given the defining characteristics of most dementia syndromes are primarily cognitive in nature, assessment of a person's cognition is a vital component of both diagnostic services and research investigations, and is the most common outcome measure by which the effectiveness of potential pharmaceutical and non-pharmaceutical therapies is judged. Standardised paper-and-pencil cognitive assessment tools are a key component of the screening and diagnostic process, but have a number of limitations. Accurate assessments are long and associated with participant fatigue and stress, but brief tests often elicit floor and ceiling effects owing to a lack of dynamic range [1]. Literacy and education effects on cognitive scores due to the high linguistic demands of instructions, lack of reproducibility due to the assessor's subjectivity bias and ecological validity of some cognitive domains (e.g. social cognition) are further potential confounding factors [2].

Recent studies suggest that eye-tracking-based cognitive assessment might ameliorate some of the existing problems as it enables a brief and quantitative evaluation of cognitive functions [2]–[4]. Eye-tracking technology provides fine-grained information regarding oculomotor information (pupil dilation and gaze) and has been used to uncover eye movement abnormalities in different dementia syndromes [4], [5]. Previous studies explored its usability mainly for diagnostic purposes using it as a proxy to cognition during basic oculomotor functions (e.g. saccadic behaviour) and for evaluation of particular higher-order cognitive functions (e.g. memory, attention) [6], [7]. Recently Oyama et al. [2] used it as a communication tool during cognitive assessment to collect answers from patients with dementia and mild cognitive impairment that indicated their preference with their gaze while the tasks instructions were written on the screen. Although these tests capture critical aspects of task performance, they are still susceptible to the need for instructing patients on how to complete the tasks, which is prone to mistakes caused by misunderstandings, language

K. Mengoudi, D. Ravi, D. Alexander are with Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK. (e-mail: kyriaki.mengoudi.16@ucl.ac.uk).

S. Crutch, K. Yong, E. Brotherhood, K. Lu, J. Schott and I. Pavisic are with Dementia Research Centre, Queen Square Institute of Neurology, Department of Neurodegenerative disease, University College London, London, UK.

I. Pavisic is also with UK Dementia Research Institute at University College London

S. Primativo is with the Department of Human Science, LUMSA University, Rome, Italy.

difficulties or patients at the later stages of the disease. Novel instruction-less tests might be a window to more natural, robust and ecologically valid cognitive evaluation.

Currently, the most commonly used way of summarising eye-tracking information is the computation of statistics over the pupil dilation and the gaze signal. The latter is converted to a sequential series of events predominantly consisting of fixations (eyes held stable), saccades (rapid movements to change the position of fixations) and blinks. Because of the high level of variability between individuals, a single measure such as mean fixation duration per spatial unit (e.g. a word in a text) is not able to capture characteristics relevant to cognitive processes and thus a set of features are calculated on carefully selected spatial areas of interest. Other methods for eye movement analysis include statistics and heatmaps over raw gaze data, similarity indices of scanpaths, as well as, the so-called n-grams features that encode information for the direction and the amplitude of eye movements [8]–[10]. In dementia research, the previous features take the form of abnormalities and are expressed in terms of latency, accuracy, stability and variability [11]. However, the identification of a complete set of handcrafted features from cognitive tests sensitive to subtle task and participant-specific abnormalities is non-trivial and time-consuming. Additionally, these features are not generalisable to more complex stimuli because they rely on specific stimulus characteristics (e.g. regions of interest).

To overcome the limitations of handcrafted features, researchers have explored different computational approaches using unsupervised representation learning; by learning an embedding that captures some of the semantics of the input placing semantically similar inputs close together in the embedding space [12]. Self-supervised representation learning is a promising subclass of unsupervised representation learning which has produced state-of-the-art visual representations in standard computer vision problems [13]. This method uses information already present in the data as a supervision signal so that supervised learning techniques can be used. The rationale behind self-supervised learning is that by training a network to solve a pretext task, it encodes high-level semantic representations that are useful for solving other tasks of interest that usually have little annotated data. For sensors data, supervised representation learning with deep learning models has been shown to be competent in tasks including Human Activity Recognition (HAR) from wearable devices and detection of seizures or arrhythmia from electroencephalogram (EEG) and electrocardiogram (ECG), respectively [14]–[16]. However to our knowledge for eye-tracking data, supervised representation learning has only been used for detection of gaze events (e.g. fixations, saccades) from raw eye-tracking sequences and self-supervised representation learning has not been exploited [17].

In this work, we introduce a novel way of detecting abnormal behaviour and automatically extracting salient features from a novel instruction-less eye-tracking cognitive test administered to well-characterised patients with a variety of dementia diagnoses and healthy controls. We use the pretext task of identifying the particular cognitive task from which a particular eye-tracking sequence came. Labels are well-defined and known from this task and it supports self-supervised learning to identify salient features of eye-tracking sequences. Our results not only validate that instruction-less eye-tracking tests can detect dementia status but also reveal novel self-supervised learning features that are more sensitive than handcrafted features in detecting performance differences between participants with and without dementia across a variety of tasks.

II. Materials

A. Datasets

Controls A: Eye-movement data from 432 healthy adults between 18 and 82 years were collected during a residency at the London Science Museum as part of the C-PLACID project. Thirty-one of these (mean age: 62.03 [SD: 7.79], 19 females [F], 12 males [M]) were over fifty years old, had proficient skills in English and reported no neurological conditions, visual impairment or dyslexia.

Controls B: Data from the Insight 46, a sub-study of the National Survey of Health and Development (NSHD) (British 1946 Birth Cohort) were also used for validation. 144 healthy individuals (67 F : 77 M) born in the same week in 1946 underwent the eye-tracking test and standard cognitive assessments at age 69-71 years. 121 of these individuals were cognitively healthy and amyloid negative based on Amyloid PET imaging.

Patients: Thirty patients with dementia (10 F : 20 M) participated in the study with mean age 68.9 years (SD : 9.16), of which 20 were less than 65 years of age at the time of their diagnosis. In terms of disease severity, their average MMSE score was 22.6 (SD: 6.68) and 18 of the patients had mild symptoms (based on correspondence with Clinical Dementia Rating scale; MMSE$>$20) [18]. These participants fulfilled standard clinical criteria for diagnosis of one of the following dementia subtypes: AD (6 subjects), bvFTD (7), lvPPA (5), nfvPPA (6) and svPPA (6).

B. Stimuli and Procedure

All participants (patients and Controls A & B) completed a free-viewing eye-tracking test in which 48 images were presented on a computer screen for 3 seconds each (for a total of 192s) and their eye movements were recorded using a desk-mounted video-based eye-tracker (Eyelink 1000 Plus) at 1000 Hz. Participants were not given explicit task instructions; they were just asked to look at the screen. A chin rest was used to maintain a constant viewing distance of 80cm in all participants. Stimuli were selected to engage different cognitive functions:

1) Scene exploration: i) social interaction; 10 images with social (people present) and non-social context (people absent) (e.g. Figure 1 a.) , ii) missing items; 10 images,

Fig. 1: Example stimuli from the five cognitive tasks illustrated as they were presented on the computer screen sequentially (one image at a time) in the order administered: a. social interaction; image with people present (social), b. missing items; a chair with a missing part, c. social scenes; combination of pictures depicting two gardens scenes with a person present on the right, d. semantic processing; an example of a semantically incongruent sentence; "She likes having a cup of injury in the morning", e. recognition memory; the previously presented picture (from the social interaction task) on the right (see a.) is coupled with a new picture on the left.

half complete and half incomplete (e.g. Figure 1 b.) and iii) social scenes; 8 images depicting either a garden or a kitchen scene where a person is present on one side of the screen and absent on the other (e.g. Figure 1 c.).

2) Semantic processing: 10 sentences, half of which were semantically congruent (e.g. "In the jungle there are many different animals.") and half semantically incongruent (e.g. "She likes having a cup of injury in the morning."), administered in pseudorandom order (e.g. Figure 1 d.).
3) Recognition memory: 10 pairs of images, one of which was seen previously in the social interaction task and the other which is a new image of equivalent style and complexity (e.g. Figure 1 d.).

There were four different versions of this test. V1-2 included all tasks (semantic processing, scene exploration, recognition memory) but had different stimulus sets. V3-4 included the same stimuli as V1-2 but excluded the semantic processing task. Controls A (Science Museum) participants were randomly assigned to one of the four versions. Controls B (Insight 46), Controls A elderly and patients were administered either version V1 or 2.

The eye-tracker was calibrated for each participant using 9 calibration points. Each trial was initiated by the experimenter and every trial was preceded by a centrally presented fixation point used as a drift correct stimulus. The fixation point also enabled a drift check, as the experiment only proceeded if the participants was looking at the drift target. Images were presented in a fixed random order within each task, and tasks were administered to all participants in the same order.

III. Methodology

To mine the information of the eye-tracking time series of this instruction-less eye-tracking test, we implemented the following steps (Figure 2):

I) Cognitive activity recognition: Firstly, self-supervised representation learning was implemented in which condensed abstract representations of the input signal are learnt training a deep neural network on Cognitive Activity Recognition (CAR) based on healthy individual's data (Controls A).
II) Feature relevance visualisation: Once the distribution of healthy behaviour was learnt, LRP was used to explain the networks' decisions in differentiating between cognitive activities and eventually to better understand the mechanisms underpinning healthy behaviour.
III) Abnormality detection: Next, the extent to which eye-tracking features of dementia patients deviate from healthy behaviour was explored.
IV) Dementia classification: This was followed by a comparison between self-supervised and handcrafted representations on discriminating between participants with and without dementia.

For the following analysis, the models' performance was evaluated in terms of $F1$ score which is the harmonic mean of precision and recall.

A. Data Processing

The EyeLink system recorded gaze position and pupil size in a monocular tracking mode providing 1000 samples per second. Gaze position reports the (x, y) coordinates of a subject's gaze on the display (resolution: 1920 x 1080) in actual display coordinates (pixels) with origin (0, 0) at the top left. Pupil size is reported as the pupil area measured in arbitrary units typically ranging between 100 to 10000 units. Raw samples, therefore, consist of three-time series of x, y coordinates of gaze and pupil size having a dimension of [sampling rate x trial duration].

Eye movement events were generated by the EyeLink tracker including fixations, saccades and blinks using standard velocity and acceleration thresholds. Saccades identified as containing blinks were considered blinks. Trials with total number of samples outside the screen's resolution or total blink duration more than 500ms were considered erroneous and were excluded from the analysis.

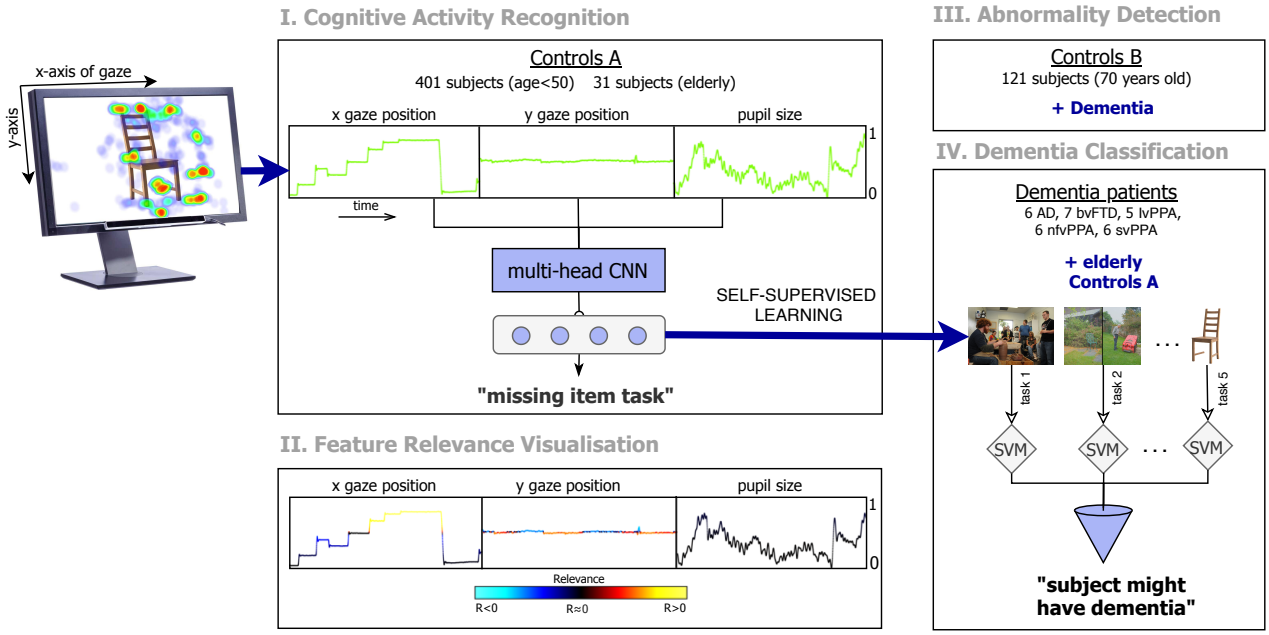

Fig. 2: Outline of the Methodology: I. Two multi-head Convolutional Neural Networks (CNN), model A and B, were trained to identify the particular cognitive task from which a particular eye-tracking sequence came based on healthy individual's data (Controls A). II. Heatmaps based on Layer-wise Relevance Propagation technique applied on model A were visualised that show areas of the input that particularly contribute to a prediction of a cognitive task. III. Controls B and Dementia data were fed to model B for trial and subject-wise abnormality detection. IV. Model's B features learnt through self-supervised learning (by training initially a deep neural network to recognise distinct cognitive activities in healthy individuals) were transferred for subject-wise dementia classification using an support vector machine majority voting scheme.

Gaze position signal was normalised to the display coordinates by dividing the gaze coordinates by the screen resolution. Missing values of gaze position were imputed with a constant zero value to avoid interpolation bias; as missing values might have a physical meaning indicating fatigue or cognitive load.

Processing of pupil size data involved discarding data before and after blinks and linear interpolation of missing values and lowpass Butterworth filter with cut-off frequency of 5 Hz. This cut-off frequency was found to be optimal for noise minimisation and signal restitution in our data. The baseline pupil size was measured as the average pupil size for a period of 300 ms immediately preceding each stimulus onset. This baseline value was selected because firstly it is a duration long enough to give a robust estimate which is longer than the average blink duration. Secondly, it is small enough to minimise the influence of pupil dilations from a previous trial since the inter-trial intervals in the battery are 1000 ms. Baseline corrected pupil diameters were computed by subtracting the baseline pupil size from the raw pupil size after stimulus onset.

1) Handcrafted Features: The following basic eye movement statistics were chosen to summarise the free-viewing tasks of the experiment:

saccade counts, total duration of saccades, median of the length of saccades (x-coordinate of gaze), number of progressive saccades (forwards), number of regressive saccades (backwards), fixation counts, mean/max/standard deviation, blinks counts, total duration of blinks and total duration of fixation duration, mean/std/min/max of peak velocity, visual angle, pupil size, pupil size during fixations, x and y coordinates of gaze.

Scanpath length, namely, the Euclidean distance of saccadic movements with respect to x and y position of gaze, was also selected as a measure of the overall functional performance of participants since it has been associated with higher fluid intelligence scores in healthy individuals [19].

Overall differences in eye-movement handcrafted features across all tasks between healthy controls and dementia patients were evaluated using a Generalised Estimating Equation (GEE) model with independence correlation structure and robust standard errors to adjust for repeated measures for each subject [20]. In addition to the group category (controls/dementia), the following variables were included in the GEE models: age, education, gender, task and task by group interactions.

B. Representation Learning Methodology

1) Cognitive Activity Recognition: Cognitive activity recognition from eye movements was used in this work as the pretext task and a deep neural network was trained in a trial-wise manner (rather than subject-wise) given the raw eye-tracking signals. CAR can be considered as a classification problem, where the inputs are time series and the out-

puts are the cognitive task one is being assessed on (semantic processing; scene exploration; recognition memory). In particular, a multi-head CNN architecture was implemented which takes as an input the three eye-tracking time series, processes them separately by individual one-dimensional convolutional heads and extracts features specific to each time series [21]. This network's architecture processes the entire sequence at once generating a single feature map for each sample and then all the features maps are concatenated. In this way, the features extracted from each time series are kept separated which improves the interpretability of the model and captures better data of different natures and scales that are not correlated (e.g. gaze coordinates and pupil size). After the feature extraction stage, a global average pooling operation was applied which calculates the average output of each feature map and prepares the model for the final classification layer.

Defining the number of output classes is not straightforward, as the instruction-less nature of the test increases between-trial variability causing label ambiguity. Although the stimuli were designed to trigger specific reactions, it is not guaranteed that participants were performing in a similar/uniform way, especially on the missing item, social scenes and interaction tasks which fall broadly into the scene exploration task.
The following two multi-class problems were investigated which differ on the number of output classes in the final classification layer:

i) Model A: Three-class problem (scene exploration, recognition memory and semantic processing task)
ii) Model B: Five-class problem (missing items, social scenes, social interaction, recognition memory and semantic processing task)

Figure 3 demonstrates the architectures of model A and B. Model A consists of a single 1-D convolutional layer (kernel size 5, stride equal to 1, no padding) with 5 features maps for each input signal followed by a batch normalisation layer and a ReLU activation function. Model B includes 12 blocks of the following architecture in the order presented: 1-D convolutional layer (kernel size 5, stride equal to 1, no padding) with 30 features maps, batch normalisation layer, ReLU activation function, dropout layer ($p = 0.2$) and average pooling layer (pool size 2). In both cases, a global average pooling layer follows the feature extraction block and reduces the dimension to 15 and 90 features for model A and B respectively. These features are the input of a perceptron applied with a softmax activation function.

2) Training Details: Hyper-parameter selection and model comparisons were implemented within the following pipeline: data were split in train and test set under the constraint that trials of an individual appear in only one of the sets. 5-fold cross-validation was implemented on the train set for each combination of parameters selected using grid or random search. The set of parameters with the best 5-fold cross-validation score ($F1$ score) was selected and the model with weights from the best fold was evaluated on the test set.

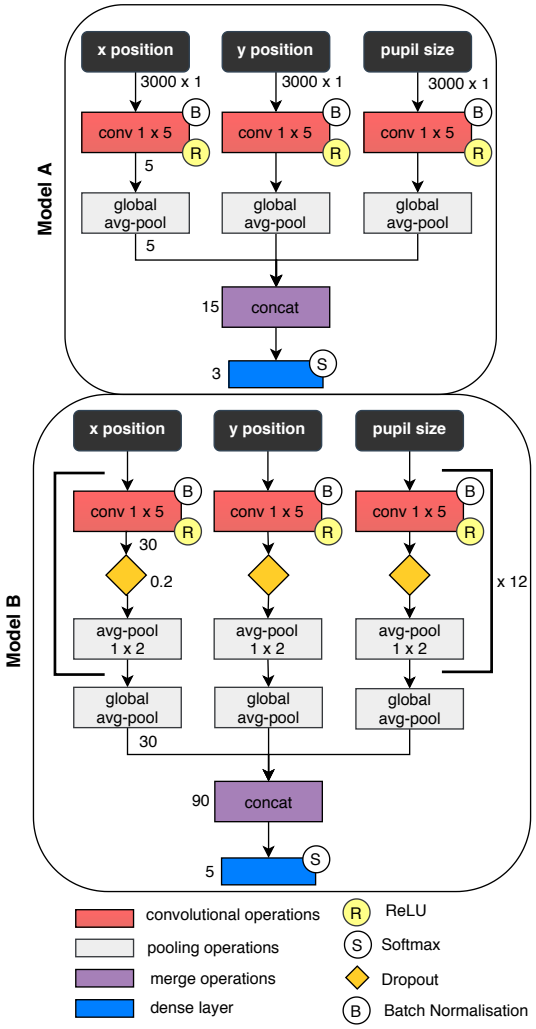

Fig. 3: Model A and B Neural Networks Architectures for cognitive activity recognition with 3 (scene exploration, recognition memory and semantic processing task) and 5 (missing items, social scenes, social interaction, recognition memory and semantic processing task) output classes respectively.

The proposed framework was implemented and trained in KERAS. The network parameters were optimised by minimising the categorical cross-entropy loss function using gradient descent with Adam optimiser having learning rate of 0.001 and batch size of 50. Weights were randomly orthogonally initialised with the Glorot normal initialiser. The L1-norm weight regularisation was applied with regularisation rate of 10^{-5}. The maximum number of epochs was 50 and early stopping was implemented which stops the training process if the validation loss does not increase for 50 contiguous epochs. After training the model, only the weights of the epoch with the higher $F1$ score on the validation set were saved and used for the evaluation of the model.

3) Data Augmentation: Given the intrinsic within and between person variability and the limited amount of eye movement data, data augmentation can be used to prevent overfitting and improve the generalisability of the models [22]. Finding invariant properties of the data against certain transformations is the main idea behind the selection of the following four techniques implemented: shifting, jittering, scaling and cropping. For each gaze position time series, transformations were applied by randomly selecting two out of the four techniques.

Shifting involves generating samples by shifting x and y coordinate of gaze by a scalar (randomly sampled from the interval [-10, 100] for the semantic processing task and from [-100, 100] for all other tasks). In this way, we covered unexplored input space by accounting for variability of the movement while preserving the shape of it. Jittering is a way of simulating additive noise attributed to varying levels of gaze stability in individuals or noise associated with the sensor. Three seconds of Gaussian noise was generated with a standard deviation value sampled from a uniform distribution $U(0.05,\ 1)$. Scaling the input by multiplying the x and y coordinates of gaze with the same scaling factor attempts to change the magnitude of the signal and subsequently slightly the shape of the gaze scanpath. The scaling factor was sampled from the normal distribution with a mean of one and standard deviation between 0.05 and 0.2 for the semantic processing task or 0.1 for the others. Cropping the input involves removing the first (or last) x time points (x in [5, 100]), shifting the signal x points in the time axis and consequently interpolating with zero values to keep the original dimension.

C. Feature Relevance Visualisation

LRP, proposed in [23], was applied to the best performing CAR model to better understand the mechanisms underpinning healthy behaviour during different cognitive activities. LRP attempts to explain the decisions of non-linear models such as deep neural networks. The goal of this technique is to quantify the contribution of each component of an input a to the prediction $f(a)$ made by a given decision function f. To this aim, LRP decomposes f attributing relevance scores R_i to all components i of a such that $f(a) = \sum_i R_i$. The algorithm starts from an output neuron j by defining $f(a) = R_j$ and it iterates over all the layers of the model backwards to the input attributing relevance messages to each neuron under the constraint that the total amount of relevance is conserved in each layer.
The relevance value being propagated from neuron j to its input i is proportional to each input i contribution to the activation of the neuron j and is defined as:

$$R_{i\leftarrow j} = \frac{z_{ij}}{z_j} R_j, \tag{1}$$

where where z_{ij} is the contribution of the input neuron i to the output neuron j and $z_j = \sum_i z_{ij}$. In this work a modification of this formula is used, the so-called ϵ-rule, which introduces a stabiliser $\epsilon > 0$ to the denominator of formula 1 to avoid possible unbounded values of $R_{i\rightarrow j}$ with small values of z_j. The relevance score R_i at input neuron i is then obtained by summing all incoming relevance values $R_{i\rightarrow j}$ from the output neurons to which i contributes to and is defined as:

$$R_i = \sum_j R_{i\leftarrow j}. \tag{2}$$

By replacing $R_{i\leftarrow j}$ with the above formula, it is obvious that a neuron is relevant if it contributes to neurons that are relevant themselves.

In the CAR classification setting, for input neurons i, $R_i \approx 0$ indicates inputs with no or little influence on the model's decision, $R_i > 0$ represents parts of the input that explain a specific class while $R_i < 0$ contradicts the prediction of that class.

Once the relevance values were computed, they were normalised to the interval $[-1, 1]$ by dividing with the maximum absolute relevance value of the entire input signal.

D. Abnormality Detection

Once the normal behaviour during this cognitive assessment was learnt, the extent to which the eye movement patterns of dementia patients deviate from it was investigated. In particular, a question of interest is whether dementia patients passively look at the screen without following the implied instructions (e.g. reading when a sentence is presented on the screen) which is the expected activity from the controls. To investigate that, data from unseen elder controls (controls B) and dementia patients were fed into the pre-trained CAR neural network and the number of misclassified samples were estimated for each cognitive task and group. Since misclassifications might be attributed to behaviour patterns not seen in the training set, abnormality was defined in relation to an unseen elderly controls' dataset. A threshold that discriminates normal from abnormal cases for each cognitive task was created by calculating the average predicted probability of an elder control trial belonging to each cognitive task classes. A trial from the dementia group was considered abnormal if the model assigned it to a class with probability less than the threshold value of that specific class. Finally, a majority voting strategy was applied to determine abnormal participants of the dementia cohort using the median value of the abnormality scores of their trials.

E. Dementia Classification

Since the ultimate purpose of cognitive assessment is the detection of dementia related oculomotor biomarkers, we evaluated whether the representations learnt using the cognitive activity recognition task (i.e. pretext task) are useful for dementia classification (i.e. target task). If the representations learnt are general and not specific to the pretext task, then the target task is expected to perform well. To this aim, the data of the elderly healthy controls A and dementia

patients were fed to the pre-trained CAR neural network and the outputs of the average global pooling layer were the features to be transferred for dementia classification. The performance of a support vector machine (SVM) classifier with these abstract features and handcrafted features was compared. The following procedure was applied to both feature sets.

The features were adjusted by controlling for potential confounding effects of gender, age and education levels before being fed to the classifier. A multivariate multiple linear regression model was fitted on the controls' data with the features as dependent variables and age, education and gender as independent variables. Subsequently, the residuals were calculated for the features matrix which measure unexplained variance presumably attributed to the task or to other individual characteristics. Using the model with the estimated coefficients the residuals were also calculated for the patients' features. The inputs, therefore, of the classifier were the residuals instead of the initially calculated features. In addition, all the features were standardised by removing the mean and scaling to unit variance inside the cross-validation procedure using the mean and variance of the train set.

For each cognitive task, we extracted abstract or handcrafted features for each trial and then made trial-wise predictions of dementia status based on all set of features, whether abstract or handcrafted. Five SVMs with a radial basis function (RBF) and tuning parameters the kernel coefficient (γ) and the penalty parameter (C) were fitted to the features of each task separately (missing items, social scenes, social interaction, recognition memory and semantic processing task) [24]. This ensemble approach was preferred to a single global classifier, because we hypothesised that the task information will improve the predictions. Lastly, to obtain subject-specific from trial-wise predictions, a majority voting scheme (median operation) was applied to the five classifiers' predictions; twice for each subject (Figure 2). In more detail, for each cognitive task, the corresponding SVM made several predictions (votes) for all trials of each subject (e.g. for semantic processing 10 predictions for each subject). The final prediction for each subject's performance on a particular task was the one that received the most votes. Finally, the global output prediction of each subject was the one that received more than 3 votes (out of 5).

Nested cross-validation was implemented for the evaluation of the classifiers: data were split into a train set, within which parameters were selected with 5-fold cross-validation (γ in $[10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}]$ and C in $[0.001, 0.01, 0.1, 1, 5, 10, 25, 50, 100, 1000, 1500]$), and a test set, for evaluation. It was ensured in the process that the same participants appear in the test sets for all five classifiers. This process was repeated 100 times and since the classifier's performance was evaluated in terms of $F1$ score, 100 $F1$ scores were obtained for each experiment.

A label permutation test was implemented as a baseline which determines the performance of the model when there is no relationship between the features and the output labels. The features of the best performing model were selected for this procedure. Comparisons between the performance of the model with different feature sets were made using Mann-Whitney U test and bootstrap confidence intervals with 1000 iterations were calculated.

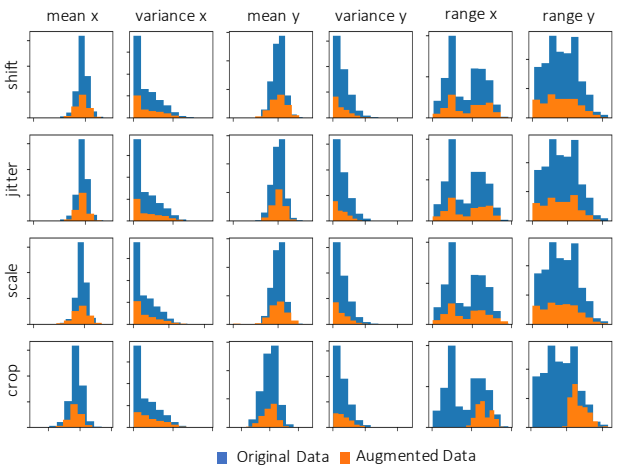

Fig. 4: Histogram of statistics for original and augmented data. Statistics: mean, variance and range of the time series signal were computed for all samples for x and y coordinate of gaze.

IV. Results

A. Cognitive Activity Recognition

Table I summarises the results of the CAR model, which classifies cognitive activity given eye-tracking data, in terms of 5-fold cross-validation and left-out test set performance with two combinations of output classes (3 or 5 output neurons) and four combinations of training datasets (with or without augmented data, controls B and combined controls A and B). The original dataset includes 15,996 trials and the augmented 18,793. Figure 4 shows that the distribution of the computed statistics (mean, variance, range) for x and y coordinates of gaze for original and augmented data are similar. In cropping, the range of the samples is higher because the minimum value is always zero, as the signal is interpolated. The best performance in terms of $F1$ score on the test set appears to be on the simplest model with three classes (scene exploration, memory and semantic processing) trained on the original dataset of healthy controls. Data augmentation and increasing the size of the training set (A and B) improves slightly the performance of CAR5 but not CAR3 model.

B. Feature Relevance Visualisation

Relevance maps were computed for the best performing model (CAR3) for both dominantly and not-dominantly firing output neurons as the latter can reveal interesting information about the learnt strategy of the model e.g. why a certain class has not been picked for prediction. Figure 5

Table I: Performance scores of the multi-head CNN models on activity recognition with different multi-class and augmentation settings evaluating with 5-fold cross validation and the left-out test set.

Model	Control Dataset	Output classes	Augment	CV $F1$	Test $F1$
CAR3	A	3	False	0.955 (0.01)	0.967
CAR3_AUG	A	3	True	0.948 (0.014)	0.954
CAR3	B	3	False	0.941 (0.014)	0.926
CAR3	A + B	3	False	0.946 (0.012)	0.959
CAR5	A	5	False	0.841 (0.023)	0.821
CAR5_AUG	A	5	True	0.857 (0.019)	0.834
CAR5	B	5	False	0.854 (0.014)	0.859
CAR5	A + B	5	False	0.852 (0.01)	0.854

provides some insight on the methods the network uses to classify with high certainty eye-tracking trials belonging to the scene exploration (neuron 1), semantic processing (neuron 2) or memory recognition task (neuron 3). It shows the contribution of the input to the prediction of each class, or in other words, to the output of each neuron of the final layer of the model. Positive values of relevance in the not-dominantly firing neurons indicate parts of the input sharing properties with the dominant neuron. Negative values in the not-dominantly firing neurons indicate parts of the input that significantly oppose the properties of the dominant neuron.

Figure 5.a constitutes an example of a semantic processing trial of a healthy control which is correctly classified with 0.982 probability. The neural network attributes positive relevance to peaks, high values or gradually increasing stair-wise trends of the x coordinate of gaze. The network discriminates between scene exploration and semantic processing or recognition memory largely by these x position properties of the signal. This is reflected in the negative relevance of the same points in neuron 1 compared to neuron 2 and 3.

Figure 5.b displays an example of a memory recognition task (class 3) which is correctly classified by the network with probability 0.981. As seen in Figure 5.a previously, positive relevance values are attributed primarily in the x position of gaze in both neuron 2 and 3. Here to discriminate between memory recognition and semantic processing, the network looks at the y position of gaze and gives higher values of relevance to fixations (flat areas) with higher values (bigger jumps) of the y coordinate of gaze. Intuitively this means that it learns that in the semantic relative to the memory task the eyes stay relatively still with respect to the vertical axis of the screen while moving horizontally to read the sentence.

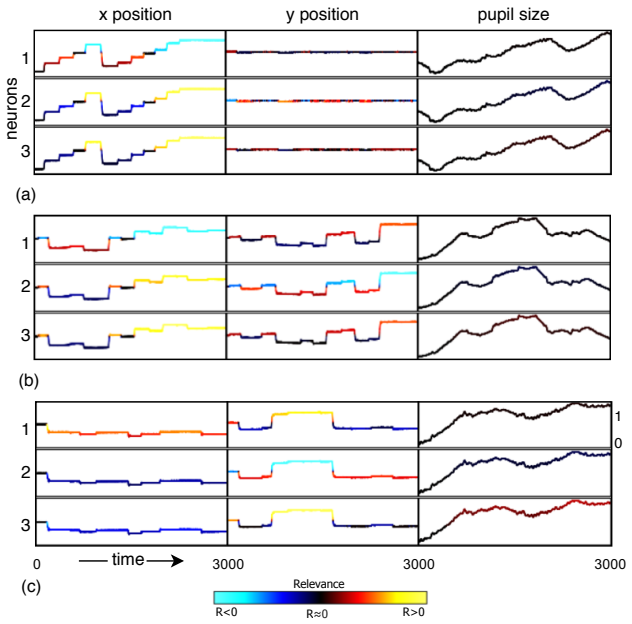

Fig. 5: Relevance plots of Cognitive Activity Recognition (CAR) features discriminating between cognitive tasks in healthy controls. Features are presented from the best performing activity recognition model (CAR3). Three different input eye-tracking samples (a, b, c) are presented, each of a separate healthy control performing a reading (a) or episodic memory (b, c) task. Rows: Relevance maps with respect to the network's (class representing) output neurons (neuron 1: scene perception, neuron 2: reading, neuron 3: episodic memory). Columns: x, y coordinate of gaze and pupil size respectively. Warm hues (standing for $R > 0$) identify input components supporting the model prediction and cold hues (mapped from $R < 0$) pointing out evidence in the input considered as contradictory to the learned class by the model. (a). The model indicates that peaks, high values or gradually increasing stair-wise trends of x coordinate of gaze are associated with either the reading or episodic memory task. (b). These trends accompanied by relatively stable values of y position of gaze are attributed to the reading task, whereas long fixations with higher values (bigger jumps) of the y coordinate of gaze to the episodic memory task. (c) In samples where there are no jumps with respect to the horizontal axis of the screen, the network identifies big jumps in y position of gaze and variations in pupil size as features associated with the episodic memory task.

In the memory recognition task when the previously seen feature of x position of gaze is not apparent, i.e. there are no jumps with respect to the horizontal axis of the screen, the network looks for big jumps at the y position of gaze (Figure 5. c, probability = 0.99). Interestingly, since this property is shared between scene exploration and memory recognition task, the network classifies the trial as memory recognition relying also on the pupil size signal.

C. Abnormality Detection

Based on the results of the handcrafted features, overall dementia patients searched less extensively and scanned the stimuli significantly more slowly than controls with lower scanpath lengths (z = -276.56, SE = 97.09, z = 8.11, p = 0.00439). Mean x position of gaze was lower in dementia patients compared to controls (z = -22.895, SE = 11.220, z = 4.16, p = 0.0413). There was also a significant interaction between the effects of group and task ($p < 0.0001$); dementia patients showed a greater relative impairment relative to controls in the semantic processing task, looking at lower values of the x coordinate of gaze on average when the sentences appeared on the screen. The same patterns appear on median coordinate of gaze (z = -26.07, SE = 10.34, z = 6.35, p = 0.012).

To investigate whether dementia patients passively look at the screen without following the implied instructions, the CAR5_AUG model trained on healthy behaviour was used. The percentage of misclassified trials when the controls B validation set vs dementia data were fed into the model were higher for the dementia patients for all the cognitive tasks: social scenes (Controls: 27.4% vs Dementia group: 37.3%), semantic processing (0.4% vs 2.7%), memory recognition (5.7% vs 8.8%), social interaction (22.4% vs 28.5%), missing items (13.6% vs 19.2%). The distribution of the predicted probabilities of a trial belonging to a task were statistically significantly different between controls B and dementia patients trials apart from the social interaction task (social interaction: z = 20067.5, p = 0.067, semantic processing: z = 24236, $p < 0.0001$, missing items: z = 14576.5, p= 0.0084, social scenes: z = 13682, $p < 0.0001$, memory recognition: z = 17305, p = 0.0002).

In terms of the detection of abnormal participants, even in the absence of explicit task instructions, 13 out of 30 dementia patients were considered abnormal in the social scenes task (threshold p = 0.6851), 10 in the social interaction task (p = 0.71) and 4 in the missing items task (p = 0.808).

D. Dementia Classification

Figure 6 and Table II summarise the results of the model on the dementia classification task using handcrafted and deep learning features obtained from different variations of the CAR models presented above. Overall, the features from CAR5 present the best results capturing differences between the two groups (95% CI: [0.7870, 0.8241]). The handcrafted features [0.6175, 0.6723] show lower performance compared to CAR5_AUG [0.7522, 0.7944] (t = 2334 , $p < 0.0001$), CAR5 (t = 1628, $p < 0.0001$) but not CAR3_AUG [0.6412, 0.71097] (t = 4371, p = 0.094) and CAR3 [0.6367, 0.6979] (t = 4287, p = 0.064), mainly as CAR3_AUG and CAR3 are performing better, but this improvement is not statistically significant. There was no evidence that data augmentation improved classification performance in the CAR5, (t = 4195.5, p = 0.013) nor in CAR3 problem (t = 4894.5, p = 0.398). Both handcrafted features and CAR5 differ significantly from the baseline case (t = 3479, $p < 0.001$, t = 1628, $p < 0.001$).

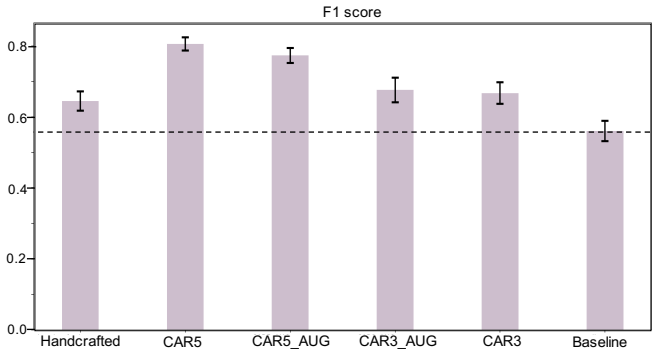

Fig. 6: Performance of the SVM-ensemble model trained on the dementia classification task in terms of $F1$ score using different handcrafted and deep learning features. The self-supervised learning features were transferred from different variations of the CAR neural network trained on activity recognition. CAR3 and CAR3_AUG have three output neurons (scene exploration, recognition memory and semantic processing task) and are trained with and without data augmentation, respectively. CAR5 and CAR5_AUG have five output neurons (missing items, social scenes, social interaction, recognition memory and semantic processing task). For a baseline, the case where there is no relationship between the features and the output labels is considered. Bars represent 95% bootstrap confidence intervals.

Table II: Mean of 100 iterations of nested cross-validation metrics (TN: True Negative, FP: False Positive, FN: False Negative, TP: True Positive) of the SVM-ensemble model trained on the dementia classification task and tested on 6 patients and 6 controls.

Model	TN (%)	FP (%)	FN (%)	TP (%)
CAR3	37.08	12.91	16.33	33.66
CAR3_AUG	33.5	16.5	15.25	34.75
CAR5	33.5	16.5	5.16	44.8
CAR5_AUG	30.5	19.5	6.33	43.66
Baseline	11.08	38.91	12.91	37.08
Handcrafted	34	15.98	17.94	32.05

V. Conclusion

A mixed sample of well-characterised dementia patients varying in disease severity participated in a free-viewing test which was designed to assess specific cognitive functions selectively impaired in different subtypes of dementia. In this study, we present evidence that firstly brief instruction-less eye-tracking tests can detect abnormal oculomotor biomarkers and secondly self-supervised representation learning techniques can extract more discriminative features from this instruction-less eye-tracking cognitive test that are more discriminative than standard handcrafted eye-tracking metrics.

To assess the overall functional performance of participants in the cognitive test, scanpath length and some relevant handcrafted features were computed. We found that dementia patients search less extensively and scan the stimuli significantly more slowly than controls. They also present a tendency to fixate towards the left side of the screen during sentence presentation compared to controls, which might indicate that either are slower in reading or they are not reading the sentences. While these primary findings are unable to indicate the basis of such abnormal performance, such as whether this relates to a diminished ability to adapt eye behaviour in response to task demands [5], they demonstrate that features extracted from even this brief and instruction-less test may detect abnormal oculomotor biomarkers of dementia-related cognitive dysfunction.

With the aim to evaluate whether the dementia patients were performing the activities that were implied by the test, we first tried to understand the cognitive behaviour of the average healthy control. We trained a neural network on healthy controls to predict cognitive activity from eye movements and the network's decision strategies were analysed. Its decisions are determined by different channels of the input (e.g. x position of gaze) associating the combination of jumps towards high values of x position and stable y position or jumps towards y position with the sentence reading task (since one normally scans the screen from left to right when reading) and the recognition memory respectively. Interestingly, to discriminate between scene exploration and memory recognition, the networks seem to use information from pupil size which is consistent with previous evidence of pupil response being modulated during memory tasks [25].

When this network was used to classify cognitive activities of elderly controls and dementia patients, higher misclassification errors were observed in dementia patients than in controls indicating that dementia patients perform the distinct cognitive tasks differently than healthy participants. If slower performance in a task is associated with eye-tracking sequences with the same features to the average performance but shifted later in time, then we know that trials of slower participants are not misclassified by the network. This is because according to the equivariance to translation property of convolutional neural networks, two different input signals with the same feature presented in different locations in the input space produce the same output. Misclassified trials of dementia patients, therefore, might be attributed to cases where the mechanisms underpinning cognitive activities differ substantially to controls.

The cognitive activity recognition pretext task not only contributes to the detection of abnormal behaviour but also provides general condensed representations of the eye-tracking data useful for dementia classification from a variety of cognitive tasks. This is demonstrated by the ability of our framework to predict dementia status in this heterogeneous group with an $F1$ score between 0.7870 and 0.8241. This result was achieved with abstract features obtained from the most complex model (deep neural network) trained on the most difficult classification problem (activity recognition of five cognitive tasks) with 90 features in total. Although this model achieves a significantly lower performance on activity recognition than other less complex models, it learns richer representations of eye-tracking data that are more sensitive in detecting performance differences between participants with and without dementia. In addition, all sets of abstract features outperform standard handcrafted features, highlighting the added value of new feature extraction techniques for eye-tracking data from cognitive tests especially under the lack of instructions. These findings demonstrate the importance of self-supervised representation learning to healthcare applications in the absence of a large number of patients data.

To the best of our knowledge, this is the first application of deep learning for classifying and interpreting cognitive activity and dementia status from raw eye-tracking measurements. These methods were applied to a particularly complex dataset that included different versions of an instruction-less cognitive test with varying levels of stimulus complexity (abstract scene viewing versus simple sentence stimuli). Additionally, the test was also administered to clinically well-characterised patients, not only those with typical presentations, but a combination of rare dementia syndromes varying in disease severity. Our results show that self-supervised representation learning methods hold promise for augmenting cognitive assessment with instruction-less eye-tracking tests to monitor patients at different stages of the disease in a brief, low-stress manner.

Although this study opens the door to a more ecologically valid assessment of natural cognitive behaviour in dementia, more work needs to be done. The current method is not able to show whether the features are sensitive to the different dementia subtypes nor evaluate the effectiveness of the specific parts of the tests to the targeted groups (e.g. memory test for tAD patients). This can be potentially addressed in the future with the recruitment of larger within-subtype dementia cohorts. Future evaluation of patients in the early stages of dementia, with mild cognitive impairment or living at autosomal dominant genetic risk of a dementia, might determine whether this battery can be used for early detection of cognitive change/impairment. In addition, from a methodological perspective, although the current dementia classification task shows whether the features learnt in the pretext task are meaningful for dementia-related abnormality detection, it might not be the best approach for screening patients highlighting abnormalities in different subtypes and stages of the disease. The reason being is that it assumes a homogenous pattern of abnormality in the dementia group, which might not be true given the variability of eye movement behaviour between subjects. Anomaly detection based on detecting outliers given a distribution of normal behaviour might be a more appropriate tool here for future research.

To conclude, this work highlights the contribution of self-supervised representation learning techniques in medical applications where the small number of patients, the non-homogenous presentations of the disease and the complexity

of the setting can be a challenge using state-of-the-art methods. It also demonstrates that the application of methods for interpreting artificial intelligence systems constitutes a window to better understand human cognitive functions. The proposed methodology of the unsupervised representation learning technique with the LRP interpretability framework presented above is applicable to different cognitive tests, instruction-less or not, under the only assumption that they include activities associated with distinct eye-movements.

ACKNOWLEDGMENT

This work was supported by a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 666992. It was also supported by EPSRC grants EP/M020533/1 and EP/M006093/1 and the NIHR UCLH Biomedical Research Centre. Insight 46 was principally funded through grants from Alzheimer's Research UK (ARUK-PG2014–1946, ARUK-PG2017–1946), Medical Research Council, Dementias Platform UK and the Wolfson Foundation. Keir Yong is funded by the Alzheimer's Society, grant number 453 (AS-JF-18-003).

REFERENCES

[1] M. C. Morris, D. A. Evans, L. E. Hebert, and J. L. Bienias, "Methodological issues in the study of cognitive decline. American journal of epidemiology,," *American journal of epidemiology,*, vol. 149, no. 9, pp. 789–793, 1999.

[2] A. Oyama, S. Takeda, Y. Ito, T. Nakajima, Y. Takami, Y. Takeya, T. Katayama, H. Rakugi, and R. Morishita, "Novel Method for Rapid Assessment of Cognitive Impairment Using High- Performance Eye-Tracking Technology," *Scientific Reports*, vol. 9, no. 1, pp. 1–9, 2019.

[3] A. P. A. Bueno, J. R. Sato, and M. Hornberger, "Eye tracking – The overlooked method to measure cognition in neurodegeneration?" *Neuropsychologia*, p. 107191, 2019.

[4] S. Primativo, C. Clark, K. X. X. Yong, N. C. Firth, J. Nicholas, D. Alexander, J. D. Warren, J. D. Rohrer, and S. J. Crutch, "Eyetracking metrics reveal impaired spatial anticipation in behavioural variant frontotemporal dementia," *Neuropsychologia*, vol. 106, pp. 328–340, 2017.

[5] T. J. Shakespeare, Y. Pertzov, K. X. X. Yong, J. Nicholas, and S. J. Crutch, "Reduced modulation of scanpaths in response to task demands in posterior cortical atrophy," *Neuropsychologia*, vol. 68, pp. 190–200, 2015.

[6] T. J. Anderson and M. R. Macaskill, "Eye movements in patients with neurodegenerative disorders," *Nature Reviews Neurology*, vol. 9, no. February, 2013.

[7] R. J. Molitor, P. C. Ko, and B. A. Ally, "Eye Movements in Alzheimer ' s Disease," *Journal of Alzheimer's Disease*, vol. 44, no. 1, pp. 1–12, 2015.

[8] S. Hoppe, T. Loetscher, S. A. Morey, and A. Bulling, "Eye Movements During Everyday Behavior Predict Personality Traits," *Frontiers in Human Neuroscience*, vol. 12, p. 105, 2018.

[9] A. Bulling, S. Member, J. A. Ward, H. Gellersen, and G. Tr, "Eye Movement Analysis for Activity Recognition Using Electrooculography," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 741–753, 2010.

[10] S. K. Mannan, K. Christopher, and M. Husain, "The role of visual salience in directing eye movements in visual object agnosia," *Current biology*, vol. 19, no. 6, pp. R247–R248, 2009.

[11] I. M. Pavisic, N. C. Firth, S. Parsons, D. M. Rego, T. J. Shakespeare, K. X. Yong, C. F. Slattery, R. W. Paterson, A. J. Foulkes, K. Macpherson, A. M. Carton, D. C. Alexander, J. Shawe-Taylor, N. C. Fox, J. M. Schott, S. J. Crutch, and S. Primativo, "Eyetracking metrics in young onset alzheimer's disease: A Window into cognitive visual functions," *Frontiers in Neurology*, vol. 8, pp. 1–16, 2017.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning : A Review and New Perspectives," vol. 35, no. 8, pp. 1798–1828, 2013.

[13] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation Learning by Learning to Count," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

[14] O. Faust, Y. Hagiwara, T. Jen, O. Shu, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals : A review," *Computer Methods and Programs in Biomedicine*, vol. 161, pp. 1–13, 2018.

[15] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G. Z. Yang, "Deep Learning for Health Informatics," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[16] D. Rav, C. Wong, B. Lo, and G.-z. Yang, "Deep Learning for Human Activity Recognition : A Resource Efficient Implementation on Low-Power Devices," *In 2016 IEEE 13th international conference on wearable and implantable body sensor networks (BSN)*, pp. 71–76, 2016.

[17] R. Zemblys, D. C. Niehorster, and K. Holmqvist, "gazeNet : End-to-end eye-movement event detection with deep neural networks," *Behavior Research Methods*, vol. 51, no. 2, pp. 840–864, 2019.

[18] K. Komossa, T. Grimmer, J. Diehl, and A. Kurz, "Mapping Scores Onto Stages : Mini-Mental State Examination and Clinical Dementia Rating," *American Journal of Geriatric Psychiatry*, vol. 14, no. 2, pp. 139–144, 2006. [Online]. Available: http://dx.doi.org/10.1097/01.JGP.0000192478.82189.a8

[19] B. A. Sargezeh, A. Ayatollahi, and M. Reza, "Investigation of eye movement pattern parameters of individuals with different fluid intelligence," *Experimental Brain Research*, vol. 237, no. 1, pp. 15–28, 2019.

[20] B. Y. K.-y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[21] M. Canizo, I. Triguero, A. Conde, and E. Onieva, "Multi-head CNN – RNN for multi-time series anomaly detection : An industrial case study," *Neurocomputing*, vol. 363, pp. 246–260, 2019.

[22] T. T. Um, F. M. J. Pfister, L. München, D. Pichler, S. Endo, M. Lang, U. Fietzek, and D. Kulić, "Data Augmentation of Wearable Sensor Data for Parkinson' s Disease Monitoring using Convolutional Neural Networks," *arXiv preprint arXiv:1706.00527*, 2017.

[23] S. Bach, A. Binder, G. Montavon, and F. Klauschen, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[24] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 20, pp. 273–297, 1995.

[25] A. Kafkas and D. Montaldi, "The pupillary response discriminates between subjective and objective familiarity and novelty," vol. 52, pp. 1305–1316, 2015.