Computational and Structural Biotechnology Journal 24 (2024) 264–280

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

EM-COGLOAD: An investigation into age and cognitive load detection using eye tracking and deep learning

Gabriella Miles [a,*], Melvyn Smith [a], Nancy Zook [b], Wenhao Zhang [a]

[a] Centre for Machine Vision, Bristol Robotics Laboratory, University of the West of England, T Block, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, UK
[b] Faculty of Health and Applied Sciences, University of the West of England, Bristol BS16 1QY, UK

ARTICLE INFO

Keywords:
Time series classification
Eye movement
Deep learning
Cognitive load
Age

MSC:
0000
1111

PACS:
0000
1111
2000

ABSTRACT

Alzheimer's Disease is the most prevalent neurodegenerative disease, and is a leading cause of disability among the elderly. Eye movement behaviour demonstrates potential as a non-invasive biomarker for Alzheimer's Disease, with changes detectable at an early stage after initial onset. This paper introduces a new publicly available dataset: EM-COGLOAD (available at https://osf.io/zjtdq/, DOI: 10.17605/OSF.IO/ZJTDQ). A dual-task paradigm was used to create effects of declined cognitive performance in 75 healthy adults as they carried out visual tracking tasks. Their eye movement was recorded, and time series classification of the extracted eye movement traces was explored using a range of deep learning techniques. The results of this showed that convolutional neural networks were able to achieve an accuracy of 87.5% when distinguishing between eye movement under low and high cognitive load, and 76% when distinguishing between the oldest and youngest age groups.

1. Introduction

Alzheimer's Disease (AD), first documented by Alois Alzheimer in 1906, is the most prevalent neurodegenerative disease (ND), accounting for 60–75% of dementia cases [1,2]. AD progressively affects memory, cognitive faculties, and behaviour. Initially asymptomatic, progression is marked by cognitive and functional decline, with impairment and neuropsychiatric symptoms becoming increasingly severe as the disease progresses [3].

The diagnosis of AD involves a comprehensive evaluation of the individual, necessitating multiple steps [4] in a complex process, frequently requiring input from several healthcare professionals. The typical route to diagnosis begins with general practitioners (GPs), who are the first point of contact for patients and their families. GPs conduct initial clinical tests, and if cognitive decline is suspected, refer individuals to specialists, such as neurologists and psychiatrists [4]. Neurologists perform neurological examinations and neuroimaging, and assist in differential diagnosis - distinguishing AD from Parkinson's Disease (PD), for example [4]. Psychiatrists carry out more in-depth cognitive or mental status testing, as well as reviewing patient history, often working closely with neurologists [4,5]. Given the length of this process, and the fact that observable symptoms[1] - such as personality changes - frequently only manifest long after the initial onset of the disease, timely diagnosis poses a challenge. The timely diagnosis of AD (and NDs generally) is an area in need of improvement as highlighted by the Institute of Medicine [6]. In particular, the early diagnosis of AD is linked to the potential attenuation of cognitive decline [7], which has significant impact on both the individual with AD and their families.

The current AD diagnosis approach is effective but time-consuming, involving input from trained professionals across various tests aimed at detecting skill or functional loss, which are challenging to detect in the early stages. The application of artificial intelligence (AI) for assessing cognitive function through eye movements potentially offers a user-friendly, passive evaluation method suitable for various settings like GP surgeries, clinics, or homes. Results might be instantly provided to clinicians in a clear and relatable format. Given the projected increase of global AD cases [9], there is a need for an assessment tool that can be used at the early stage, and across a large population.

* Corresponding author.
E-mail address: gabriella.miles@bristol.ac.uk (G. Miles).

[1] It is important to note however, that individuals in the preclinical stage of AD may exhibit impaired working memory and executive function at an early stage in the disease development, even in the absence of other observable symptoms [8].

https://doi.org/10.1016/j.csbj.2024.03.014
Received 29 November 2023; Received in revised form 15 March 2024; Accepted 16 March 2024
Available online 27 March 2024

In recent years, changes in eye movement have continued to show potential as a non-invasive biomarker for NDs, including for mild cognitive impairment (MCI), and AD. In addition to changes due to ND, eye movement characteristics also change across the human lifetime. For example, research has shown that the reaction time (RT) of ballistic eye movements, such as saccades, typically slows [10–13], and that destination targeting becomes less accurate [14] with age, even among healthy older adults. Smooth pursuit eye movements (SPEM) also demonstrate an increased initiation RT [15,16] and increased saccade frequency [16]. These age-related differences are further exacerbated by the presence of AD or MCI [17] on both pro- [18] and anti-saccade [18, 19] as well as SPEM [20,21] tasks.

While eye movement analysis has been investigated for the purpose of distinguishing between normal cognitive function, MCI, and AD [22–25], conventional eye tracking methods (typically employing head-mounted eye trackers) often involve a large amount of pre-processing, such as segmenting eye movement time series data into individual saccades or fixations and studying them in isolation. This paper instead focuses on the application of deep learning (DL) techniques to the analysis of long-sequence, non-periodic eye movement data (>30, 000 frames per sample), with a specific emphasis on healthy and atypical ageing.

To achieve this, the authors' conducted a study which recruited healthy individuals. These individuals had self-reported to be cognitively unimpaired, and had not been diagnosed with any of the exclusionary criteria (detailed in Appendix D). Participants performed a visual tracking task and a mental arithmetic task - while their eye movement was recorded with a single desk-mounted camera, as well as a short cognitive test measuring inhibitory control (the Simon task).

Previous research has shown that both working memory (WM) [26] capacity and cognitive load can have a significant effect on eye movement [27–29]. As such, two test conditions were created with the intention to alter cognitive load involved in each task: low cognitive load (LCL) and high cognitive load (HCL). In the HCL condition, participants carried out the LCL (visual tracking) task while simultaneously performing the mental arithmetic task. These varying cognitive load conditions were employed with the aim of (i) creating cognitive variability (using a dual-task approach similar to that undertaken in [30, 31]), and (ii) exploring their effect on eye movement and whether the proposed machine learning (ML) approach is sensitive enough to differentially detect such variability.

The purpose of this research was to validate the use of DL techniques in detecting eye movement changes resulting from a lack of cognitive resources. This would demonstrate potential for the early detection of MCI and AD, which is characterised by an overall reduction in WM [32], using DL. The application of DL techniques in this context may result in more objective (data-driven), and earlier interventions. This is of significant importance given that early diagnosis can influence the prognosis of the disease, with indications suggesting that the rate of cognitive decline may possibly be reduced [7]. As such, there is increasing demand for automated tests capable of determining cognitive ability - the research presented in [24] has been developed into a tool for simple and rapid cognitive assessment [33], with the aim of providing differential diagnoses of dementia. This tool offers the ability to obtain objective assessments, which could be adopted as a costeffective solution for clinical use. Our approach differs from that described in [24], in which the duration that users spent looking at specific regions of interest is measured and correlated to the cognitive score. In the methods used in this paper, the full eye movement trace is analysed and minimal pre-processing of the data is required, permitting the investigation of subtle and complex patterns in eye movement data which may elude traditional analytical methods.

A head-mounted alternative [34] to that presented in [33] aims to diagnose many different NDs, while [35] also employs eye tracking technologies but focuses on the diagnosis of Attention Deficit Hyperactivity Disorder (ADHD) and dyslexia.

Traditional assessments to diagnose dementia can often be stressful for the individual, and can also require travel to specialist clinics. By interacting with a visual display, rather than actively engaging with questioning or challenging tasks, the stress response of participants may be reduced. The eye movement tests described in this paper are 3–4 minutes long, and require minimal instructions. Given the reduced workload (compared to more traditional methods of assessment) of applying and analysing such tests, they could be used to obtain continuous data on cognitive function, potentially facilitating early intervention, as well as the evaluation of such care.

The remaining structure of this paper is as follows. The experimental procedure is described in detail in Section 2.1, while the datasets collected are detailed in Section 2.2. The method used to extract time series data from the dataset is outlined in Section 2.3, with the results of the eye state detection (ESD) and eye centre localisation (ECL) models detailed in Sections 3.1 and 3.2, respectively. Methods for the analysis of this time series data are described in Section 2.4, with corresponding results in Section 3.3. Results for the Simon task are detailed in Section 3.4. Finally, the study and results are discussed in Section 4, with concluding remarks given in Section 5.

2. Material and methods

This section describes the experimental protocol (Section 2.1), the gathered dataset (Section 2.2), the method for extraction of the eye movement time series data (Section 2.3), and the methods for analysing the data to distinguish between cognitive load conditions and age groups (Section 2.4).

2.1. Data capture experiment

The aim of this experiment was to elicit different types of eye movement under varying cognitive loads and capture images of the participants' eyes as they watched the designed videos. By increasing the cognitive load associated with a task, we aim to emulate the challenges faced by those experiencing cognitive decline [30,31,36] - characterised by a reduction in working memory capacity. The images were used to train DL models for the purpose of extracting key features pertaining to eye movement. The context was to investigate the potential for ML/DL techniques to distinguish between cognitive decline in typical and atypical ageing. Section 2.1 details the experimental procedure (Section 2.1.1) and participant demographic information (Section 2.1.2). Further information regarding the specific constituent elements of Video 1 (V1) and Video 2 (V2) are given in Appendix A and Appendix B, respectively.

2.1.1. Experimental procedure

The experimental setup included a computer and associated peripheral accessories (mouse, keyboard etc); a forehead-chin rest to minimise participants' head movement; a 27-inch monitor, positioned at a distance of 57 cm from the participant; and a high-speed camera (FLIR Grasshopper 3) set to record at a resolution of 400 × 960 at 160 frames per second, positioned at a distance of 40 cm from the participant. The experiment took place in a quiet, well-lit laboratory setting, with additional lighting also positioned above the monitor.

Upon arriving at the experiment location, participants reviewed the participant information sheet - which had been made available to them prior. All participants then signed a consent form and completed a demographics questionnaire, which gathered data relating to age group, gender, ethnicity, educational attainment, and exclusionary criteria. Participants were made aware of their right to withdraw from the experiment, and were given the opportunity to take as many breaks as necessary between each experimental stage.

The experiment required participants to watch V1 and V2 twice, under two different cognitive loads, and to complete the Simon task [37] (a short test investigating inhibitory control) under two conditions

(using two or four colours), further described in Appendix C. The duration of the experiment, including responding to questionnaires, was approximately 60 minutes.

The Simon task is a well-established cognitive task used in psychological experiments to assess cognitive performance, particularly relating to attention and response inhibition [38,39]. In our experiment, the Simon task results provide additional ground truth to age group as general trends of increasing RT and error rate with age have been widely reported when using the Simon task [38–41]. Inclusion of the Simon task therefore serves in the validation of the DL models and interpretation of the results.

In the LCL condition, the cognitive demand was limited to that inherent to the primary visual tracking task. In the HCL condition, cognitive demand was increased by introducing a simultaneous secondary task [30,31] (counting backwards in sevens).

The experiment was split into five stages: **(i)** watch V1, LCL condition; **(ii)** watch V2, LCL condition; **(iii)** complete the Simon task; **(iv)** watch V1, HCL condition; **(v)** watch V2, HCL condition. The videos were designed to elicit different types of eye movement. Prior research has shown differences in saccadic eye movement characteristics between older and younger adults [10–14], as well as healthy older adults and AD patients [17–19]. V1 drew inspiration from these studies, focusing solely on saccadic motion. Differences have also been noted between these same groups in SPEM [15,16,20,21]. As such, V2 focused on SPEM and other patterns that required more dynamic eye movements, such as waveform tracking and figure-of-eight tests. Detailed breakdowns of the video contents are given in Appendix A and Appendix B.

The Simon task used visual stimuli, with coloured squares employed as the target stimuli (TS). The test was completed under two conditions: **(i)** with two colours and **(ii)** with four colours, where the four-colour condition placed a greater demand on working memory due to the additional rules that participants were required to remember [38]. A full breakdown of the Simon task methodology and gathered data is given in Appendix C. It is important to note that Simon task results can display considerable variability, with some older adults performing much better than others. Factors such as bilingualism, education [42,43], or lifestyle [44,45] may play a role in comparatively better performance on the Simon task at a given age, serving to preserve cognitive function.

The Simon task scores are evaluated in both the two and four colour conditions, as well as across ages, to gain an understanding of each participant's cognitive ability. This is because this may affect the accuracy of the trained DL models in determining the age group and cognitive load condition of the individual (i.e. it may be harder to distinguish an older individual who scores very well on the Simon task condition from younger individuals in the eye movement tasks).

Additionally, there is evidence that fatigue has an effect on eye movement [46]. To accommodate for variations in fatigue levels, participants were invited to take part in a repeat of the experiment described in this section at a later date. This permitted data collection under different fatigue conditions given the likelihood that identical levels of fatigue would not be experienced in each experiment. This would also introduce intra-person variability to our dataset. As the repeated experiment was exactly the same, a minimum duration of two weeks between the first and repeated experiment was required. This ensured that the practice effect - which may otherwise result in participants exhibiting improved performance on both visual tracking and cognitive tasks - was mitigated as much as possible.

2.1.2. Participant breakdown

The focus of this pilot study was on healthy ageing, and so a comprehensive list of exclusionary criteria for participants was used. This is included in Appendix D. All participants were also required to have either normal or corrected to 20/20 vision.

Overall, 75 participants (50 male, 25 female) took part in the experiment, all of whom successfully completed the entire experiment (excluding the repetition). Participants were split into four distinct age groups: 18–29 (31); 30–44 (13); 45–59 (19); and 60+ (12). In terms of ethnicity, the majority of participants identified as White (53), while the remaining participants identified as Asian (16), Black (3), or preferred not to disclose (3). With respect to education attainment, the majority of participants had obtained a degree - split into undergraduate (21), postgraduate (30), or doctorates (13), with the remaining participants having attained secondary school education (7) or alternative qualifications (2).

Of the 75 participants, 30 (22 male, 8 female) repeated the experiment. Notably, participants who repeated the experiment generally tended to be younger: 18–29 (17); 30–44 (4); 45–59 (5); and 60+ (4). The majority of repeat participants identified as White (22), while the remaining participants identified as Asian (7) or preferred not to disclose (1) this information. The educational attainment of repeat participants was: secondary school (3), undergraduate degree (11), postgraduate degree (13), doctorate (2), or other (1).

2.2. Datasets

The eye movement under differing cognitive loads (EM-COGLOAD) dataset gathered during this experiment is publicly available at [47], and is divided into four constituent parts: **(i)** labelled images, detailing the location of key eye features; **(ii)** labelled images, classifying whether the image contains a blink, or an open eye; **(iii)** the Simon task results; and **(iv)** the complete eye movement traces. Information linking each participant to a corresponding age group is also provided. The ultimate purpose of this work and developed dataset is to investigate the effect of age and varying cognitive load on eye movement.

The images captured during the experiment make up the majority of files within the dataset. The image filenames denote the time the image was saved. 40,800 images were captured during the viewing of V1, while 33,920 were captured during V2, resulting in 149,440 images per participant per experiment, as participants watched each video twice. With 75 participants taking part, this represents a substantial amount of image data. These images are stored by participant within each participant folder. V1 and V2 under the LCL condition are stored in subfolders *0* and *1*, respectively, and the images captured whilst watching V1 and V2 under the HCL condition are stored in subfolders *2* and *3* respectively.

To facilitate the training of ML models for ECL, the dataset contains hand labelled images identifying the location of the pixel coordinates for the pupil centres of the left (l_x, l_y) and right (r_x, r_y) eyes, as illustrated in Fig. 1. Given the frame rate of the camera, it was likely that consecutive images would be very similar. Thus during the labelling process every fourth image was extracted for labelling consideration and subsequently the structural similarity index measure (SSIM) [48] was used to compare and select the most different images using an experimentally determined threshold for manual labelling. This subset is composed of 13,813 images containing labelled pupil centres.

In addition, another subset of 52,506 images were manually annotated with eye state regarding its openness. This consists of 27,033 images containing blinks. There is considerable individual variability in blinking behaviour: many participants did not fully close the eye during a blink, while others may have closed one eye more than the other, or not closed one eye at all. An example of blinking behaviour when the eye is not fully closed is shown in Fig. 2. As such, a blink is defined throughout this work as the point at which the centre of the pupil is no longer visible - thus the eye can be partially open during a blink.

Blinks were identified by manually inspecting thumbnails of every fourth image to locate a blink. Once a blink was identified, images were examined on a frame-by-frame basis to determine when the pupil centre was obscured or revealed, thus defining the start and end point of the blink. All interim images were also classified as a blink. To permit the development of DL models for ESD, a further 25,473 images of open eyes in a range of positions (e.g. looking to the left, right, up, or down) and degrees of openness (e.g. partially or fully open) were also labelled.

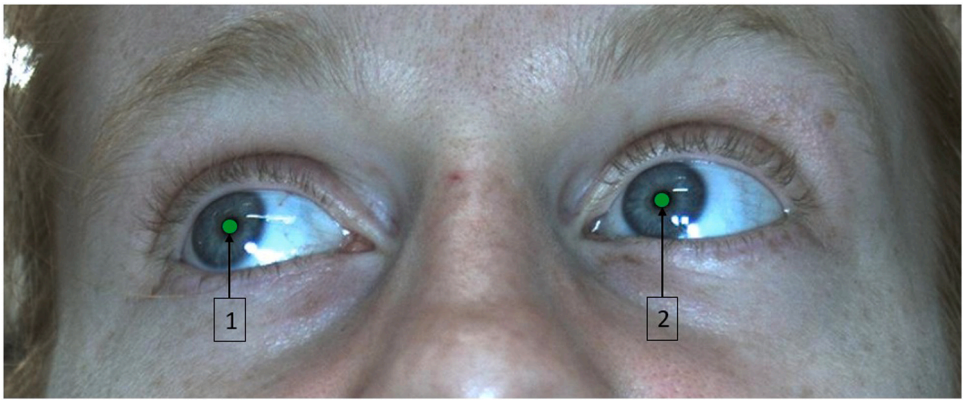

Fig. 1. Example of a labelled image, detailing the location of key eye points, resulting in four features: (l_x,l_y) and (r_x,r_y) for the left and right pupil centres, respectively. Note that *left* and *right are* denoted from the perspective of the image viewer, not the participant.

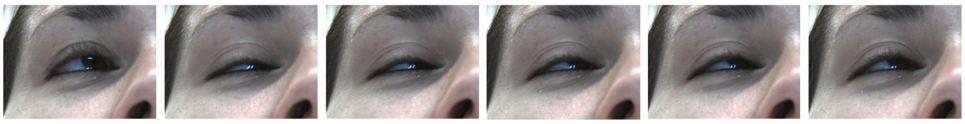

Fig. 2. An example of a partial blink cycle - illustrating the ambiguity in defining a blink. The first and last images in the sequence are defined as open eyes, while the interim images are classified as blink.

In addition to image data and corresponding labels, both the raw and calculated Simon task results are also included in the dataset and stored within the participant folders in CSV files. The raw data contains three features: **(i)** pattern, the colour and location (congruent, incongruent, neutral) of the target stimulus (TS); **(ii)** correctness, whether the participant pressed the correct key in response; and **(iii)** latency, the RT of a correct response.

The calculated data contains twenty features in total - the participant number and age group alongside eighteen calculated features: error rate, mean latency, and latency range for congruent, incongruent, and neutral trials, under LCL and HCL.

Finally, the full eye movement traces for each participant and video are also included. This data is ordered consecutively and contains five features: the *x* and *y* locations of the left and right eyes (l_x, l_y, r_x, r_y), and a binary variable indicating whether or not the image contains a blink in every captured frame.

2.3. Extraction of eye movement trace

To analyse the eye movement of the participants, it was necessary to extract the location of the eye centres in each frame. Due to the apparatus used and the position of the camera, the dataset images are constrained to a narrow field of view focused on the eye region, a subset of which is shown in Fig. 3.

While inter-participant variation is quite minimal as a result of the experimental setup, differences arise due to the presence or not of

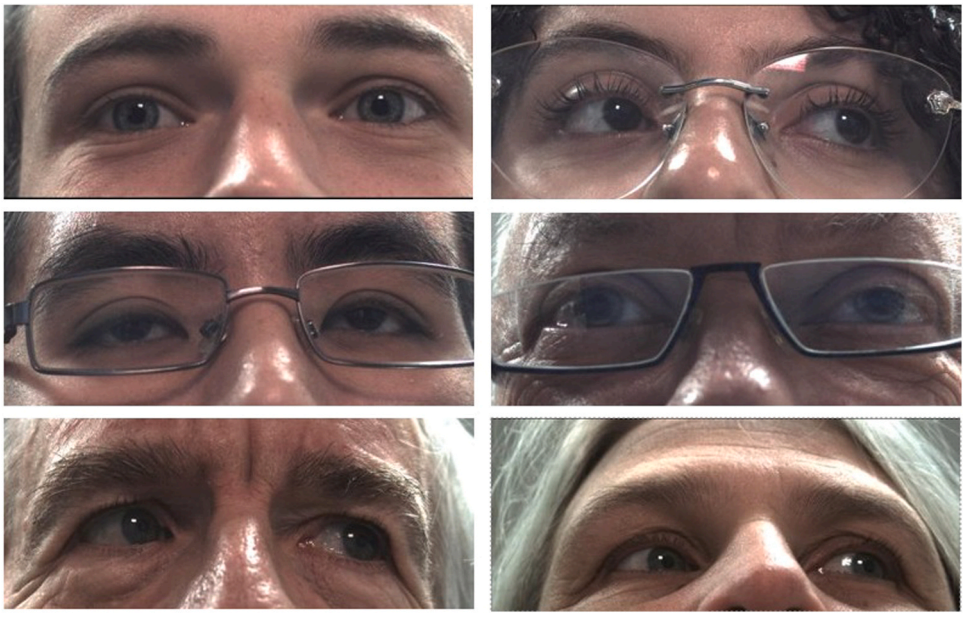

Fig. 3. Examples of dataset images demonstrating intra-person variability, as well as the overall constrained nature of the images which results in the images broadly being very similar.

glasses, as well as eye colour, ethnicity, specularities, highlights, and shadowing. Some participants also appear at more of an angle relative to the camera - which occurred when the participant did not keep their forehead in contact with the forehead bar while watching the video.

To take advantage of the large number of images captured, a DL approach was explored for the purpose of extracting eye centres. DL models, particularly convolutional neural networks (CNN)s, have shown remarkable capability to learn varied tasks from image data [49], and thus form the basis of the following investigations.

The extraction of eye centres necessitated a three-stage process composed of: **(i)** ESD; **(ii)** eye region detection; and finally **(ii)** ECL. Eye region detection was carried out as a precursor step to ECL to increase the resolution of the segmented eye region used as input to the DL models, without changing the overall size of the input image, as in [50]. Transfer learning techniques were employed using models pretrained on ImageNet [51] in each of these tasks. The pretrained models were extended by the addition of fully connected layers.

Training was a two-stage process. In the first stage, the pretrained weights were frozen, and only the added fully-connected layers were trained until the model converged. In the second stage of training, the weights of the entire model were unfrozen, including the pretrained layers. Training was carried out in this manner to prevent large updates to the pretrained model weights during backpropagation, which may result from poor initialisation of the final additional layers. Therefore, during the first stage of training the pretrained weights are used to assist in network initialisation, and in the second stage of training the model is still able to learn features that may be significantly different to the initial classification task it was originally trained on.

The pipeline used to extract the eye movement traces is shown in Fig. 4. It is composed of three stages: **(i)** ESD to identify blinks and remove these images from the remainder of the pipeline; **(ii)** individual eye region identification; and **(iii)** ECL.

The MobileNetV2 model was used for eye state detection, VGG16 was used for eye region identification, and InceptionV3 was used for ECL. All models were implemented using Keras [52] with a Tensorflow [53] backend. Results for a range of models tested are shown in Sections 3.1 and 3.2, for the ESD and ECL tasks, respectively.

During the ESD stage, all images are classified as either containing both eyes open, or a blink. For those identified as eyes open, eye region identification was carried out to locate the two patches within each image containing the eyes, and subsequently ECL to predict the pupil centres was carried out. An example of the extracted eye movement trace is shown in Figure E.12.

2.4. Determining cognitive load and age group from eye movement time series data using ML techniques

With all eye movement traces extracted according to the pipeline illustrated in Fig. 4, it becomes possible to use them to determine the cognitive load condition. This investigation was approached as a time series classification (TSC) task. The extracted eye movement time series from each video were analysed. The length of the time series is equivalent to the number of frames captured during the video. Thus the length of the V1 (saccadic motion) time series was 40,800 steps, while the length of the V2 (SPEM) time series was 33,920 steps. In each instance all four eye movement features (l_x, l_y, r_x, r_y) were used.

The objective was to classify eye movement traces as belonging to either the LCL or HCL condition. The data was split into a training set (80%) and testing set (20%). The training set was further subdivided into five folds for five-fold cross-validation, using four folds for training and the remaining fold for validation during each of the five iterations. Each participant appeared exclusively in either the training set, or the testing set. Additionally, during cross validation, participants appeared exclusively in either one of the training folds, or the validation fold. Prior to training, the eye movement traces were standardised such that they had a mean of zero and a standard deviation of one. Any sequences that were not of the requisite length were removed.

The method for determining age group from the extracted time series data was largely identical to that for cognitive load - with the same two conditions analysed, and the same four features and preprocessing steps used. However, the dataset was composed solely of participants from the youngest (18−29), and the oldest (60+) age groups, with participants once again appearing either exclusively in the training, validation, or testing sets. The objective was to classify participants as belonging either to the 18–29 or 60+ age group, thus determining the feasibility of estimating age from eye movement using DL techniques. This was to demonstrate the feasibility of discriminating between the two selected classes.

Previous research has demonstrated the efficacy of a range of models for the purpose of TSC [54,55]. The models tested in this instance were fully convolutional networks (FCN), CNN, Inception, and an encoder. The architectures of the models are shown in Appendix F, with key hyperparameter choices also detailed. All models used the Adam optimiser and the binary cross-entropy loss function. Learning rates were determined heuristically through testing on the validation sets, and varied between models.

3. Results

The results for the ESD and ECL models are given in Section 3.1, and

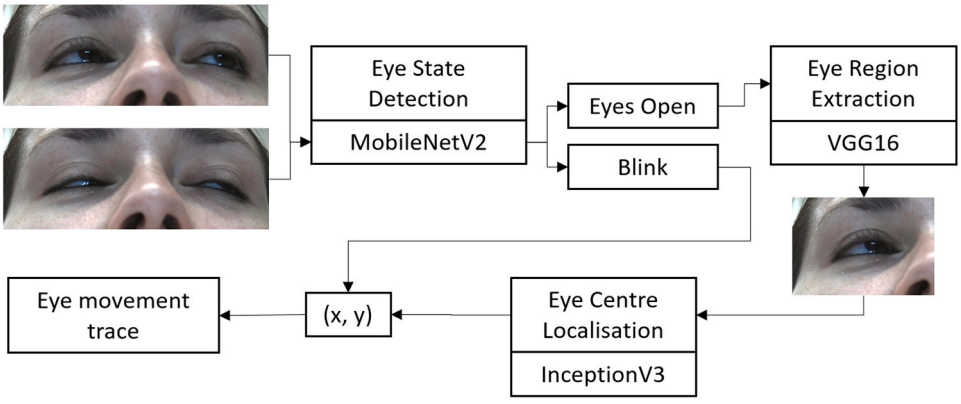

Fig. 4. Pipeline for extraction of the eye movement trace from participant images. Images are first run through the eye state detection model to be classified as a blink or not, before the eye regions of the open eyes are identified and eye centre localisation is carried out. The predicted eye centres are combined with blinks to result in the final eye movement trace.

Section 3.2, respectively. In Section 3.3 the results of the best performing DL models for determining cognitive load condition and age group from the extracted eye movement time series are detailed. Following this, a brief exploration of the Simon task results using traditional statistical techniques is given in Section 3.4 to assist with result interpretation.

3.1. Eye state detection results

The results of the ESD models on the validation set before and after unfreezing the weights of core (pretrained) model are shown in Table 1, alongside the final results of the best performing models on the test set. Three variations in the final layer configurations were explored: **(i)** the output of the core model was flattened and fed into four successive fully-connected layers; **(ii)** a global average pooling layer (GAP), which fed into a fully-connected layer; and **(iii)** a GAP layer, and two fully-connected layers. These configurations are referred to in Table 1 as 0, 1, and 2, respectively.

There was minimal performance difference between models on the validation set after the first stage of training, with each model achieving accuracies >98%. After the first-stage training step, layer configuration 1 consistently performed worse than layer configuration 2, itself performing worse than layer configuration 3 - indicating that generally, greater accuracy correlated with an increased number of parameters, although the differences in accuracy were very small (<1.35%).

While excellent results were achieved on the validation set after the first stage, the two-stage training process (in which the weights of the pretrained model were unfrozen during the second stage) generally resulted in minor performance improvements, with all models achieving an accuracy of >99%. Performance differences arising due to different model configurations were considerably reduced. The greatest improvements were seen in layer configuration 1, which had the fewest parameters. Notably, the training process demonstrated the applicability of the pretrained models to this process.

The most accurate model on the test set was MobileNetV2, with an accuracy of 99.77%, although all of the models demonstrated excellent performance. As such, processing time became a greater consideration for model selection due to the large number of images in the dataset. In addition to having the highest accuracy, MobileNetV2 processed images the fastest, and as such, was integrated into the ECL pipeline and used to detect blinks in the unlabelled images, as illustrated in Fig. 4.

3.2. Eye centre localisation results

Results for the ECL localisation on the validation set are shown in Table 2, which details the error after the first and second stages of training for the models. Normalised error, e, is reported and calculated as in Eq. 1, where d is the Euclidean distance between the ground truth and predicted centre of the left or right eyes, and w is the interpupillary distance of a given participant. Introduced by [56], it is the most widely used metric for ECL.

$$e = \frac{d}{w} \tag{1}$$

The results show that the weights of the pretrained models provide a useful starting point for training, with all three models achieving $e <= 0.1$ for >99% images in the validation set. However, there was considerable variance between models for $e <= 0.025$ after the first stage of training, with Xception and MobileNetV2 achieving this error for >70% of images. Conversely, InceptionResnetV2 only achieved this for approximately 28% and 43% of images for the left and right eyes, respectively.

The two-stage training process results in significantly improved performance for all models at the lowest error rates. The greatest improvements were demonstrated by the InceptionResnetV2 model, where the predicted pupil centre was within $e <= 0.025$ of the ground truth for 98% of images. At the smallest error ($e <= 0.025$) InceptionResnetV2 demonstrated the highest performance, as such, the performance of the model on the test set is reported in Table 3.

Table 1
Eye state detection results. Three different final layer configurations were explored for each model. Each of these configurations are identified by 0, 1, and 2, in the Layers column. TS 1 and TS 2 details the model accuracy after training stage one and two, respectively. The best layer configuration for each model was evaluated on the test set.

Model	Layers	TS 1	TS 2	Test Accuracy
Xception	0	**99.79%**	99.74%	-
	1	98.81%	**99.74%**	99.62%
	2	99.53%	99.74%	-
InceptionResnetV2	0	**99.81%**	**99.81%**	99.72%
	1	98.46%	99.76%	-
	2	99.63%	99.72%	-
MobileNetV2	0	**99.61%**	99.74%	-
	1	99.38%	**99.80%**	**99.77%**
	2	99.60%	99.77%	-

Table 2
ECL Model Accuracy for three different models on the validation set are shown for the left (L) and right (R) eyes after training stage 1 (in which the pretrained model weights were frozen) and training stage 2 (in which the weights of the entire model were unfrozen).

Model	Eye	Training Stage	$e <=$ 0.025	$e <=$ 0.05	$e <=$ 0.1	$e <=$ 0.25
Xception	L	1	72.7%	95.3%	99.8%	99.9%
		2	93.3%	**99.9%**	100%	100%
	R	1	73.8%	95.6%	99.6%	99.9%
		2	92.8%	99.2%	99.9%	99.9%
InceptionResnetV2	L	1	28.4%	76.1%	99.3%	100%
		2	**97.6%**	99.7%	99.9%	100%
	R	1	43.6%	84.2%	99.0%	99.9%
		2	**97.9%**	**99.6%**	99.9%	99.9%
MobileNetV2	L	1	75.7%	96.3%	100%	100%
		2	86.3%	98.4%	100%	100%
	R	1	75.8%	96.9%	99.8%	99.9%
		2	87.0%	98.7%	99.8%	99.9%

3.3. Eye movement time series classification analysis

Having achieved accurate ECL, the proposed DL approach proceeded to the subsequent stage, namely eye movement TSC. This section details the best performing DL models for TSC of the extracted eye movement traces for two different tasks: **(i)** distinguishing between high and low cognitive load (Section 3.3.1), and **(ii)** distinguishing between the oldest and youngest age groups (Section 3.3.2).

3.3.1. Determining cognitive load from eye movement time series results

The average results from five-fold cross-validation on the validation sets, when distinguishing between eye movement time series data under the LCL and HCL conditions alongside accuracy on the test set are shown in Table 4. The architecture of the most accurate model is shown in Table 5.

This investigation examined model performance on two different categories of eye movement time series data, which were extracted from the images of participants watching V1 (saccadic), or V2 (SPEM). The

Table 3
ECL model accuracy, for the left (L) and right (R) eyes, on the test set, using the InceptionResnetV2 model.

Eye	$e <= 0.025$	$e <= 0.05$	$e <= 0.1$	$e <= 0.25$
L	98.2%	99.5%	99.8%	99.9%
R	97.8%	99.6%	99.9%	99.9%

Table 4
Model accuracy on the validation (average scores from 5-fold cross validation) and test sets when distinguishing between LCL and HCL conditions using eye movement time series data. Saccadic indicates that only information from V1 was used, SPEM corresponds to V2.

Model	Data	Validation Accuracy	Test Accuracy
CNN	Saccadic (V1)	94.7%	**87.5%**
	SPEM (V2)	90.1%	**86.5%**
FCN	Saccadic (V1)	78.2%	-
	SPEM (V2)	70.0%	-
Encoder	Saccadic (V1)	90.1%	82.5%
	SPEM (V2)	86.2%	83.7%
Inception	Saccadic (V1)	81.1%	82.5%
	SPEM (V2)	74.1%	73.0%

Table 5
Structure of the most accurate CNN when distinguishing between cognitive load conditions.

Layer	Output Shape
Input	40800 × 4
1D Convolution	40794 × 8
1D Average Pooling	13598 × 8
1D Convolution	13586 × 64
1D Average Pooling	4530 × 64
Flatten	289920
Fully Connected	1

number of epochs to train each model was determined heuristically from the training and validation loss curves during the training and validation process and was as follows: CNN, 25; encoder, 12; and Inception, 40. The FCN model was prone to significant overfitting across a range of evaluated hyperparameters, and as such was not evaluated on the test set.

When distinguishing between HCL and LCL, the CNN model reported the best results of 87.5% on the test set. This was achieved when analysing the sacccadic data captured while participants were watching V1. Both encoder and inception models achieved equal performance, reporting 82.5% accuracy on the test set.

The CNN model also reported the best scores for distinguishing between cognitive load conditions using the V2 data, achieving an accuracy of 86.5%. In comparison, the Inception model reported considerably lower accuracy for V2 data (73.0%).

Overall, there was generally not a significant difference in performance on the test set between eye movement traces extracted from V1 or V2. This suggests that it is possible to distinguish between cognitive load conditions across a range of eye movement paradigms, particularly when using CNN or encoder based approaches.

3.3.2. *Determining age group from eye movement time series results*

The results of the TSC models when distinguishing between two age groups: the oldest (60+) and youngest (18−29), using the extracted eye movement traces are shown in Table 6. Once again, we report the average results from five-fold cross-validation on the validation sets, alongside accuracy on the test set. As previously discussed, while there is likely a general trend of cognitive decline with age it is not deterministic. Changes in eye movement behaviour are somewhat unique to the individual and their cognitive state, and thus categorising an individual as younger than they are may not necessarily constitute an incorrect classification for age. Results from the Simon task (Section 3.4) illustrate this; while cognitive ability - particularly response inhibition generally decreases with age, this is not unilaterally the case. Therefore, in future work, we propose to use the Simon task to allow for additional correlation analyses and to assist with the interpretation of results, in particular the presence of potential outliers.

Table 6
Model accuracy when distinguishing age group from eye movement traces on the validation and test sets.

Model	Data	Validation Accuracy	Test Accuracy
CNN	Saccadic (V1)	82.2%	76.2%
	SPEM (V2)	88.3%	76.9%
FCN	Saccadic (V1)	80.9%	-
	SPEM (V2)	73.1%	-
Encoder	Saccadic (V1)	76.1%	71.4%
	SPEM (V2)	73.2%	73.1%
Inception	Saccadic (V1)	46.4%	-
	SPEM (V2)	70.5%	-

The CNN and encoder models reported the best results when distinguishing between age group using either saccadic or SPEM data. Best results were achieved when analysing SPEM data captured while participants were watching V2. The FCN and Inception models generally tended to predict the majority class on the validation set, and as such performance on the test set was not evaluated.

The best performing CNN model for the age group classification task is shown in Table 7. Given the class imbalance in distinguishing between the oldest and youngest age groups, the results of the CNN model on the test set are also shown in confusion matrix form in Fig. 5 when considering saccadic data, and Fig. 6 when considering SPEM data. Note that there was a greater amount of SPEM data available once sequences of length shorter than the displayed video were removed - this occurred due to the camera not recording images during the study, or ceasing to record images while the participant was watching the video.

3.4. *Simon task results*

The Simon task results are shown in Table 8. To explore the presence of any significant interaction among age group, spatial location (congruency), and cognitive load on RT, a three-way multivariate analysis of variance (MANOVA) test was conducted using SPSS Statistics. It is important to note that MANOVA tests assume that the dependent variables within each group are normally distributed, and that the relationships between dependent and independent variables are linear - which may not always hold true.

The aim of the MANOVA was to evaluate the general performance trends with respect to age, cognitive load, and congruency, and subsequently use this to inform future work, instead of using the results solely to quantify the effect size.

The tests concerning within-subject effects revealed non-significance for the interactions of condition*congruency*age group, condition*congruency, congruency*age group, and condition*age group, suggesting the absence of interaction effects. Box's test of equality of covariance matrices yielded a non-significant outcome, indicating that the assumption of equal covariance matrices was met. For the condition*congruency interaction, Mauchly's test of sphericity indicated significance, thus Greenhouse-Geisser (GG) corrected values were used for the analysis. However, both condition and congruency exhibited significant main effects: $F(1,68) = 305.389$, $p < 0.001$ and $F(2,68) = 98.053$, $p < 0.001$, respectively. For between-subject tests, a significant age effect emerged, $F(3,68) = 8.151$, $p < 0.001$. Post-hoc pairwise

Table 7
Structure of the most accurate CNN when distinguishing between the oldest and youngest age groups.

Layer	Output Shape
Input	40800 × 4
1D Convolution	40794 × 16
1D Average Pooling	13598 × 16
1D Convolution	13586 × 64
1D Average Pooling	4528 × 64
1D Convolution	4522 × 16
1D Average Pooling	1507 × 16
Global Average Pooling	16
Fully Connected	1

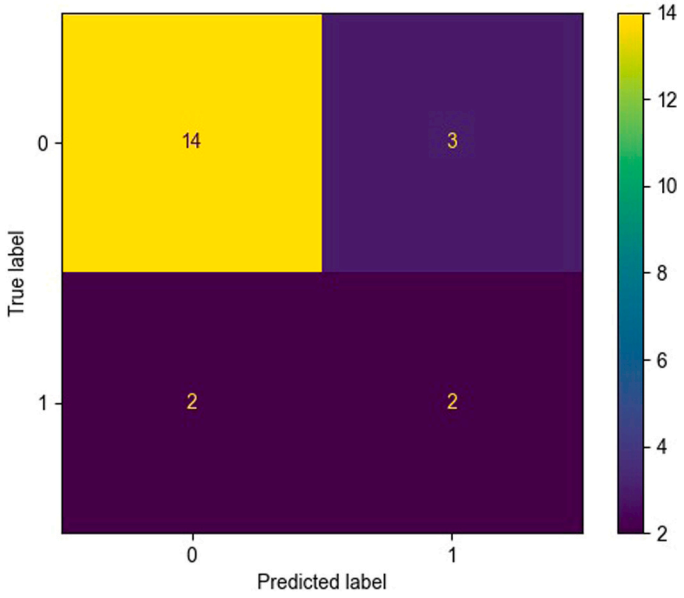

Fig. 5. Confusion matrix detailing the results of the best performing CNN model on the test set of saccadic data, showing the imbalance between oldest and youngest classes. The label of 0 corresponds to the youngest age group, while the label of 1 corresponds to the oldest age group.

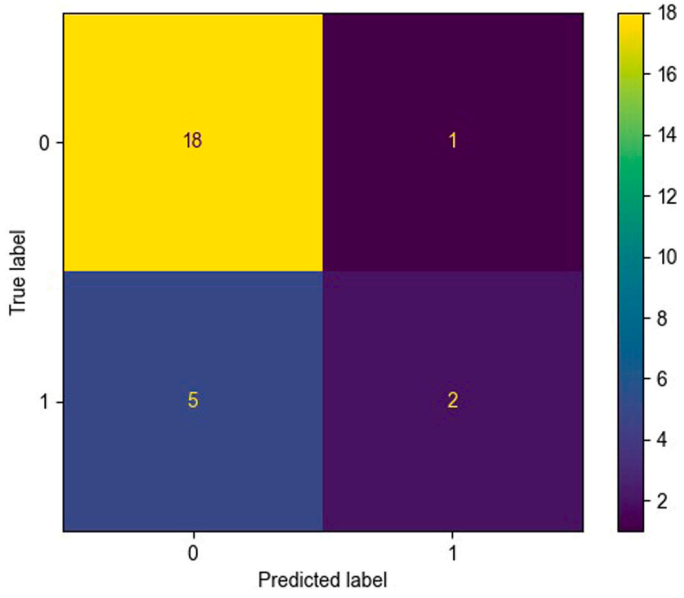

Fig. 6. Confusion matrix detailing the results of the best performing CNN model on the test set of SPEM data, showing the imbalance between oldest and youngest classes. The label of 0 corresponds to the youngest age group, while the label of 1 corresponds to the oldest age group.

comparisons employing Bonferroni corrections indicated significant differences between the 18–29 and 45–59 age groups ($p = 0.014$); the 18–29 and 60+ age groups ($p < 0.001$); and the 30–44 and 60+ age groups ($p = 0.007$), while no other interactions reached significance. Notably, discerning performance differences between the 18–29 and 30–44 age groups, and between the 45–59 and 60+ age groups was challenging, with p-values of 1.00, and 0.624, respectively.

The results of the pairwise comparisons also demonstrated significance in comparing performance between cognitive load conditions ($p < 0.001$). Furthermore, the location of the stimuli (congruent, incongruent, or neutral) exhibited significance when comparing

Table 8
Simon task RT results. Mean RT for each age group and spatial location (congruent, incongruent, neutral) in both the two and four colour condition are reported in ms.

		Age Group			
		18-29	30-44	45-59	60 +
Two-colour condition	Congruent	417.1 ± 63.8	434.1 ± 73.1	468.9 ± 72.1	494.7 ± 101.7
	Neutral	434.1 ± 74.2	455.0 ± 48.9	476.5 ± 68.4	528.8 ± 87.1
	Incongruent	459.7 ± 67.8	473.3 ± 54.6	525.8 ± 74.7	546.7 ± 75.1
Four-colour condition	Congruent	562.3 ± 66.5	561.5 ± 49.2	593.9 ± 67.0	643.1 ± 95.3
	Neutral	561.4 ± 81.1	568.1 ± 45.7	625.0 ± 67.9	657.2 ± 74.3
	Incongruent	582.5 ± 76.8	592.8 ± 56.0	650.2 ± 71.7	684.0 ± 88.5

congruent to neutral trials ($p = 0.005$), congruent to incongruent trials ($p < 0.001$), and neutral to incongruent trials ($p < 0.001$), indicating that stimuli location had a significant effect on RT. Conducting Tukey's HSD test revealed several homogeneous subsets. Notably, the 18–29 and 30–44 age groups demonstrated a very insignificant difference ($p = 0.944$), suggesting substantial difficulty in differentiating between these two age groups based on RT. The same test also found that the 30–44 and 45–59 age groups, and 45–59 and 60 + age groups did not display statistically significant differences, with p-values of 0.163 and 0.300, respectively.

Fig. 7 shows the Simon task results when the dataset is projected onto the first three principal component (PC) axes with respect to age group. From visual inspection, these age groups may not be easily separable by linear boundaries in this three-dimensional feature space (but may be more easily separable in a higher-dimensional space).

This is to be expected given the non-linear process of cognitive degeneration, as well as the considerable individual variance in performance on tests such as the Simon task due to factors such as bilingualism [38]. Despite this, the results of the PCA align with the MANOVA results, indicating much fuzzier boundaries between adjacent

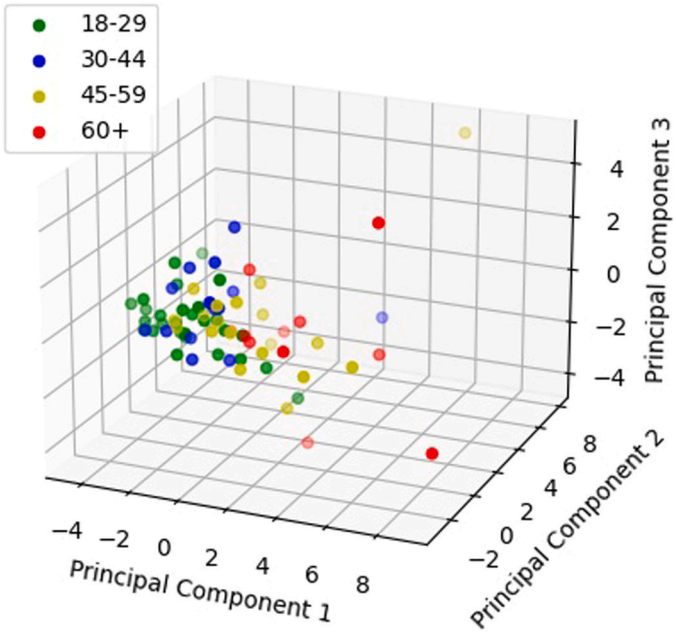

Fig. 7. Principal component analysis (PCA) results from Simon task data projected onto three PC axes, with respect to age group.

age groups (e.g. 18–29 is similar in performance to 30–44). Additionally, performance of the oldest age group appears to have the greatest variability. This could be due in part to the reduced number of participants in this age group compared to the others, or to the variability of the cognitive degeneration process. PCA was also explored for distinguishing between cognitive load (namely the two and four colour conditions in the Simon task), shown in Fig. 8.

The boundary between the two- and four-colour conditions is visually more separable than the previous distribution of age groups (Fig. 7) in this threedimensional space. This suggests that distinguishing between the high and low working memory load conditions in the Simon task may be comparatively easier than differentiating between age groups. This once again aligns with the MANOVA results, which indicate that the condition under which the Simon task was completed has a significant effect.

4. Discussion

One of the primary contributions of this work is the dataset. While the focus of this work is distinguishing between eye movement **(i)** under low and high cognitive load, and **(ii)** between the oldest and youngest age groups, the dataset contains a large number of labelled images permitting the investigation of other tasks, including ESD and ECL.

The ESD and ECL models used to extract the eye movement traces demonstrate excellent results on the test set. For the ESD task, each model achieves comparative performance, suggesting that a range of different CNN based architectures are appropriate for this task.

The efficacy of the two-stage training process employed during transfer learning is made particularly clear in the ECL task, shown in Table 2. Significantly higher accuracies were achieved after the second stage of training (in which the pretrained weights are unfrozen and subsequently updated) for each model. Thus, while the pretrained weights provide a useful initialisation for the model, further training on this specific task enhances model performance, and allows the model to specialise in learning features pertinent to ECL.

The extracted eye movement traces were used to investigate the effect of healthy ageing and cognitive load using TSC models, and a range of different models were able to distinguish between eye movement

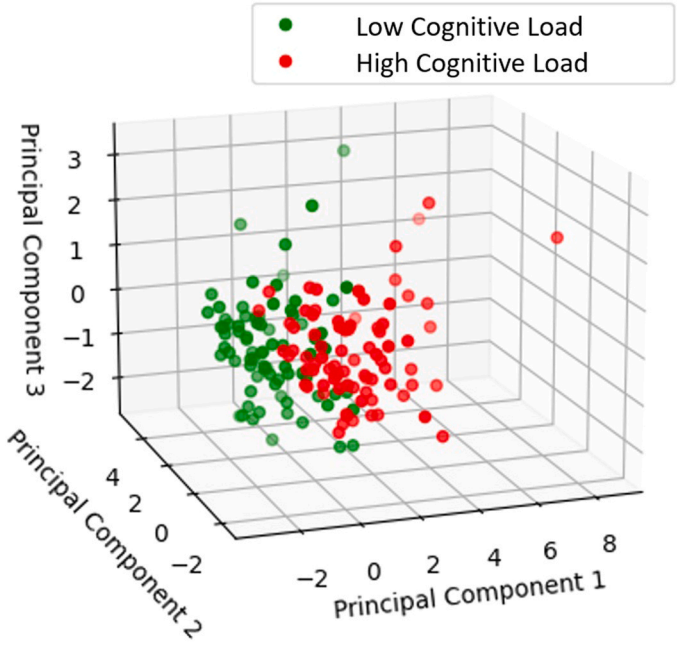

Fig. 8. PCA results from Simon task data projected onto three PC axes, with respect to cognitive load - where LCL refers to the two colour condition, and HCL refers to the four colour condition.

under LCL and HCL. CNNs in particular, were demonstrated to be effective methods for eye movement analysis, achieving accuracies of 86.5% when analysing SPEM data and 87.5% when analysing saccadic data. The comparable accuracies of the CNN model in distinguishing between LCL and HCL conditions when using saccadic and SPEM data also suggests that these differences may be detectable across a range of different stimuli. This suggests that, using our methods, changes in eye movement during day-to-day activities (or more natural visual tasks) may also be detectable. Our experiment also uses a desk-mounted camera, rather than head-mounted eye trackers. Adopting this approach has promise to enable more naturalistic eye tracking tests, especially for individuals who may be unfamiliar with head-mounted eye trackers.

A full breakdown of the CNN architectures explored is given in Appendix F. The number of convolutional layers, filters, kernel size, and ending layer configurations were investigated. The best performing model for distinguishing between cognitive load conditions is described in Table 5, which employed two convolutional layers, each followed by a pooling layer. The output of the pooling layer was then flattened, and then input to a final fully-connected layer. Due to the nature of long-sequence time series (40,800 timesteps for each saccadic trace), this resulted in a large number of parameters in the final layer. However, replacing the flatten layer with a global average pooling layer, or adding additional fully-connected layers resulted in significant performance decreases.

The addition of further convolutional layers resulted in substantial overfitting when the number of convolutional layers was greater than three. Best performance occurred when two convolutional layers were used, with accuracy highest for the (8, 16), and (8, 64) number of filters.

For the kernel size configurations, the investigations described in this paper primarily focused upon the feasibility of distinguishing between cognitive load condition and age group based on eye movement patterns that occurred in a short window. Future work will aim to better capture long-term memory in these eye movement patterns. Generally, kernel size had less impact on accuracy than number of filters, or layer configurations, although when exploring larger kernel sizes the model was more prone to overfitting.

Distinguishing between the oldest and youngest age groups was a more challenging task than distinguishing between cognitive load conditions, achieving a maximum accuracy of 76.2% on the saccadic data, and 76.9% on SPEM data using CNN models. This may stem from the comparatively smaller dataset size (75 participants for the cognitive load classification task vs 43 participants for the age group classification task), which resulted in a smaller amount of training data. Additionally, there were a greater number of training samples for the younger age group due the greater number of younger participants who took part in the repeat experiments (17 vs 5 participants). Further complexity may have been introduced to this task by using eye movement traces from both LCL and HCL conditions.

The overall accuracy when determining age group was comparable when using either the saccadic or SPEM data. However, the best performing model had a tendency to incorrectly classify older individuals as part of the younger age group when using the SPEM data. This could be due to the increased cognitive load when watching V1. The antisaccade task in particular has an inherent cognitive load greater than the saccade or SPEM tasks used in V2. Research has shown that working memory generally declines over the human lifetime, so the ability to carry out the antisaccade task (especially under the dual task paradigm) may be more impaired in older adults, making it easier to distinguish between older and younger adults - as well as resulting in more younger adults being incorrectly classified using the V1 (saccadic) data. Conversely, as the inherent cognitive load of V2 is less, the overall affect of the simultaneous task may have more measured impact on eye movement - thus making it harder to distinguish between older and younger individuals. Future work will explore the efficacy of using the individual movements designed for each video (detailed in Appendix A and Appendix B) in

distinguishing between cognitive load conditions and age groups.

Certain models (in particular Inception and the FCNs, as well as the more simplistic CNN models consisting of only two layers and minimal filters) tended towards predicting only the majority class when distinguishing between age groups. A number of approaches were applied to mitigate this - including oversampling the minority class, undersampling the majority class, and initialising the weights of the final layer with respect to the minority/majority classes, but these had minimal effect. Nonetheless, the results of the CNN and encoder models demonstrate the feasibility of distinguishing between two age group classes. In future work, methods for distinguishing between each age group (18–29, 30–44, 45–59, 60+) will be explored.

The best performing model for distinguishing between the oldest and youngest age group is detailed in Table 7. This model employed three convolutional layers, each followed by a pooling layer. The output of the final pooling layer was input to a GAP layer before this was fed to the final fully-connected output layer. As in the cognitive load TSC task, the number of filters in the layer had a greater effect on accuracy than kernel size.

PCA of the Simon task results illustrates the increased difficulty in distinguishing between age groups compared to cognitive load conditions. Despite a significant difference in performance between the 18–29 and 60 + age groups, some older adults exhibited performance that more closely matched their younger counterparts. There was also considerable overlap in the performance of adjacent age groups, notably between those aged 18–29 and 30–44.

This supports literature that indicates that cognitive decline and changes in cognitive function are somewhat unique to the individual, and dependent upon a variety of factors. When distinguishing between oldest and youngest age groups, the confusion matrix shown in Fig. 5 shows that both younger and older age groups were misclassified. Three members of the youngest age group, and two members of the oldest age group were assigned incorrect labels. In future work, eye movement will be correlated with Simon task scores, to see whether an increased inhibitory ability correlates to “younger” eye movement.

4.1. Pathway to clinical adoption

Our findings successfully demonstrate the application of CNNs for detecting changes in eye movement as a result of cognitive load and age-related cognitive decline. Thus, our study validates a DL-based, data-driven approach for the detection of small, yet significant changes in eye movement behaviour. Additionally, our approach has shown promise in detecting changes as a result of varying cognitive load and between age groups using the raw eye movement data (obtained during a short 3–4 minute video-based test) with very minimal preprocessing. This test only requires simple instructions and is easy to administer, as well as being potentially less upsetting and stressful for patients than undergoing traditional neuropyschological assessment. To the best of our knowledge, this is one of the first studies demonstrating the feasibility of distinguishing between differing cognitive loads and between age groups by applying DL to raw eye tracking measurements.

Given that both saccadic and SPEM movements were able to produce good results, eye movement data gathered during more naturalistic tasks (e.g. reading) may also be used effectively. This indicates the potential applicability of CNN models as a method for the early detection of MCI and AD using eye tracking. With AD cases rising globally, such technologies are increasingly in demand - as shown by the investment in [33–35].

It is important to note that approximately one third of GPs lack confidence in their ability to diagnose dementia [57]. Adopting these simple to administer tests in a clinical setting could therefore offer a valuable resource serving to boost confidence in diagnoses (or trigger appropriate referrals to specialist clinics), as well as having the potential to act as an early stage screening test which may alert healthcare professionals to unhealthy cognitive decline. This may lead to more targeted testing of individuals deemed at risk of cognitive decline, as well as more timely application of preventative care. In addition, opportunity exists to address the notable difference between care accessibility between urban and rural environments - which results in the pathway to dementia diagnosis being extended for individuals and their caregivers who live in rural areas [58]. The ability to administer tests remotely would help ensure that individuals are diagnosed and provided treatment in a timelier manner.

The DL models employed in this paper could provide explainable results by identifying the input sequences that support model prediction using heatmaps. This would assist in identifying changes in eye movement patterns associated with atypical cognitive decline when making predictions. An example was demonstrated by University College London (UCL) (in collaboration with the National Institute for Health and Care Research UCL Hospitals Biomedical Research Centre) [59], who developed technology to classify dementia patients and healthy controls using eye movement behaviour.

5. Conclusion

This paper has detailed the experimental procedure and gathered dataset for an experiment to investigate healthy and atypical aging by recruiting healthy participants and employing different cognitive load conditions. The dataset is comprised of several million high-resolution images, and is accompanied by hand labelled data that also permits the training of ML models for ESD and ECL. The subsequently developed ECL pipeline enabled the extraction and analysis of the participants’ full eye movement traces, which are also included for the purpose of TSC. The feasibility of recognising cognitive load condition and classifying age from eye movement data using DL models was investigated. Excellent performance was obtained for identifying the cognitive load condition, and a capability was shown for determining age group. An exploration of the Simon task results to gain a deeper understanding of cognitive ability across different cognitive loads and age groups was also carried out and showed that while age and cognitive load does influence performance on the Simon task, these changes are not universal.

While the methods and corresponding results described in this paper are at an early stage, next steps in the pathway to clinical adoption include validating the method detailed in this paper on cognitively healthy individuals compared to individuals with MCI or AD. Future work will aim to establish correlations between Simon task performance and results from the eye movement models, as well as to provide more explainable results by identifying the most salient eye movements for determining age group and cognitive load condition. Such investigations may inform the development of shorter and more specific eye movement tests. Further experiments will also be conducted, using a wider range of cognitive loads, to establish the sensitivity of the eye movement models in distinguishing between varying cognitive loads.

CRediT authorship contribution statement

Nancy Zook: Conceptualization, Methodology, Supervision. **Melvyn Smith:** Conceptualization, Resources, Supervision, Writing – review & editing. **Wenhao Zhang:** Conceptualization, Resources, Software, Supervision, Writing – review & editing. **Gabriella Miles:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC) through the FARSCOPE CDT grant EP/S021795/1. The Project Reference for this studentship is 2260377. Additional thanks must also be extended to Dr Benjamin Ward-Cherrier, Dr Alexander Smith, and Daniel Gosden for proofreading and providing exceptionally valuable feedback.

Appendix A. V1

Video 1 (V1) lasted a total duration of 255 s and focused primarily on saccadic motion. V1 employed both the gap and overlap paradigm from the gap-overlap experiment, and was split into five blocks: **(i)** prosaccades, gap condition; **(ii)** prosaccades, overlap condition; **(iii)** antisaccades, gap condition; **(iv)** antisaccades, overlap condition; and **(v)** mixed saccades, where both prosaccade and antisaccade trials were included, gap condition.

Each trial began with the display of a black central fixation cross (FC), which changed colour after 500 ms to either green, indicating an upcoming prosaccade trial, or red, indicating an antisaccade trial. In the gap condition, the FC disappeared 500 ms after the colour change, at which point a blank screen was displayed for 200 ms, before the target stimuli (TS) subsequently appeared. The TS colour always matched the FC colour, i.e. for an antisaccade trial the TS was red, and while locations were predefined the order in which they were presented was pseudorandom. This colour change is illustrated in Figure A.9.

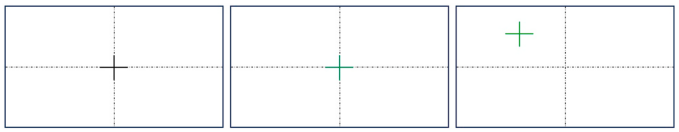

Figure A.9. An example of a saccade. In (a) the black FC is displayed, before (b) changing colour. This example illustrates an upcoming prosaccade trial (FC changes to green) - in the change of the antisaccade trial, the FC changes colour to red. Finally, as in (c) the TS is displayed, in the same colour as the FC.

.

After a further 1000 ms the TS disappeared, and the screen remained blank for 300 ms, signifying the completion of a single trial. The overlap condition differed in that the FC displayed for 1200 ms total (with the colour change still occurring at 500 ms), and the TS appeared at 1000 ms - thus both the FC and TS were visible on the screen for 200 ms from the onset of display of the TS. The TS then disappeared after 1000 ms, and the screen remained blank for 300 ms, once again marking the end of a trial. The timings for the gap and overlap saccades are summarised in Table A.9.

Table A.9
Comparison of timings for gap and overlap condition across all saccadic paradigms.

Time (ms)	Event (Gap)	Event (Overlap)
0	Fixation point appears	Fixation point appears
500	Fixation point changes colour	Fixation point changes colour
1000	Fixation point disappears	Target stimuli appears
1200	Target stimuli appears	Fixation point disappears
2000	-	Target stimuli disappears
2200	Target stimuli disappears	-
2300	-	End of trial
2500	End of trial	-

Each block consisted of 24 trials, with the exception of the mixed saccade block. The target eccentricities of the TS were ± 15°, 20°, or 24°horizontally; or ± 5°, 10°, or 15°vertically or obliquely, with each location displayed once. The mixed saccade block consisted of 48 trials, as both prosaccade and antisaccade trials were displayed at every location.

In addition to this, a calibration sequence designed to elicit consistent eye movement behaviour from participants was played at the start of the video. This was done to facilitate the temporal synchronisation of the subsequently extracted eye movement time series data across all the different trials. The calibration sequence had a duration of 14 s and consisted of a set of guided saccades in which the TS stepped from the far left side of the screen to the far right, equidistant from the top and bottom of the screen. The TS stepped every 3 s, and incorporated a countdown timer (where the colour would change every second), making the movement more predictable so that participants were able to anticipate its movement and follow the TS more accurately.

Participants watched V1 twice over the course of the experiment. The first viewing was under the LCL condition, where participants watched only the video. The second viewing was under the HCL condition, where participants watched the video while completing a mentally demanding simultaneous task - counting backwards in sevens.

Appendix B. V2

Video 2 (V2) had a total duration of 213 s and focused on SPEM, and other, less ballistic motions. The video was split into five blocks: **(i)** step-ramp tests; **(ii)** triangular waveform tests; **(iii)** random motion; **(iv)** figure-of-eight tests; and **(v)** random stepped motion. Blocks **(ii)** and (iii) also incorporated an element of predictive motion, where the TS would disappear and subsequently reappear.

Step-ramp tests, explored by [60], were designed to ensure that smooth pursuit responses are activated prior to any saccades during step-ramp tests, i.e. the eye should immediately initiate SPEM, rather than any initial catchup saccades. To achieve this, the target stimuli should *step* in a direction, and then *ramp* with constant velocity in the direction of the initial location such that it reaches its initial position approximately 200 ms after starting to move [61], as shown in Figure B.10.

Thus, the step-size during the step-ramp test is intrinsically linked to the ramp-velocity if the purpose is to elicit smooth pursuit responses from

participants prior to saccadic motion. Three ramp speeds were included in the video: 10°/s, 20°/s, and 25°/s, with corresponding step sizes of 2°, 4°, and 6°. The step-ramp tests were run in this order, with speed steadily increasing. Eight step-ramp tests were completed at a given speed, with twenty-four step-ramp tests overall. The timing and direction of the step was unpredictable, occurring 500–1500 ms after the target initially appeared at the centre of the screen. For simplicity, the sequential timings given in Table B.10 assume that the target steps at the earliest possible time (i.e. 500 ms). The total duration of the ramp was 1.0 second, after which a blank screen was displayed for 300 ms before the ensuing trial initiated by displaying the target at the centre of the screen.

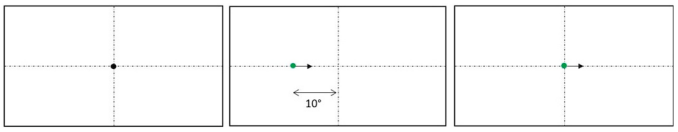

Figure B.10. (a) Central fixation point is displayed. (b) Stimulus appears at a specific distance from the fixation point and immediately begins to move in the opposite direction to which it stepped at a specific velocity. (c) The stimulus continues to move with constant velocity, until it crosses the original fixation point point 200 ms after its initial appearance.

Table B.10
Caption.

Time (ms)	Event
0	Target stimuli first appears at screen centre.
500-1500	Target stimuli *steps* and begins to *ramp*.
1500	Blank screen displays
1800	End of trial

Block **(ii)** focused on triangular waveforms. During this block each trial began with the stimulus appearing at the far left or right hand side of the screen, before it traversed to the opposite side. When the stimulus passed off screen a blank screen was displayed for 300 ms, signifying the end of the trial. In total, six trials were included for the triangular waveforms, with two trials containing sections where the stimulus disappeared for 0.5 s

The path taken by the stimulus is expressed by Equation B.1, where a is amplitude, and p is period.

$$y(x) = \frac{2a}{\pi}\arcsin\left(\sin\left(\frac{2\pi}{p}x\right)\right) \tag{B.1}$$

SPEM are most effective at tracking periodic motions that occur at frequencies of less than 0.4 Hz. Accuracy and effectiveness decreases as frequency increases to 0.5–3 Hz. As the purpose of this video was to investigate SPEM, a frequency of 0.2 was chosen for the triangular waveforms, at an amplitude of 200 pixels.

During block **(iii)** the path of the stimulus was determined using a random walk algorithm. Random motion was explored across three speeds: 10, 15, and 20 pixels/second; and also incorporated periods where the stimulus would disappear. Disappearances ranged in duration from 0.5 to 1.0 ss. Block **(iii)** was split into three stages: **(i)** speed 15 pixels/second (duration: 10 s), 2 disappearances; **(ii)** speed 10 pixels/second (duration: 5 s), 1 disappearance; **(iii)** 20 pixels/second (duration: 5 s), 1 disappearance.

Block **(iv)** featured a short section exploring eye movement behaviour during figure of eight tests. The size of the figure of eight and TS speed remained constant throughout these tests, thus forming the only periodic part of any of the videos. Five circuits of the full path were completed.

Block **(v)** was designed to mimic more natural human viewing behaviour, whilst still using a very simplistic target stimulus. This section of the video combined saccadic motion and 'random' motion aspects. Notably, this saccadic motion did not emulate the pro- and antisaccades of V1, which required participants to fixate on the FC before making a pro- or antisaccade away from this cross, returning to fixate upon it once the trial was completed. Rather, the saccadic motion in this section required participants to make consecutive saccades to the TS rather than between the TS and FC, as illustrated in Figure B.11. Overall sixty saccades were included in the stepped random motion section. The initial TS appeared towards the upper left hand corner of the screen (pixel coordinates: (100, 200)) and was displayed for one second.

Saccades were organised such that during the first portion of the section groups of smaller saccades (5–8 trials) followed by a significantly larger saccade of distance 500–800 - mimicking the switch to a different focus of visual attention. This was repeated three times, except that on the final repeat the single larger saccade was extended into a group of nine trials. TS were displayed for randomly chosen durations in the interval of 0.5–2.0 s. This entire block was then repeated again, with the TS appearing for a much shorter interval of 0.1–0.5 s

As for V1, participants watched V2 twice over the course of the experiment, under the same LCL and HCL conditions.

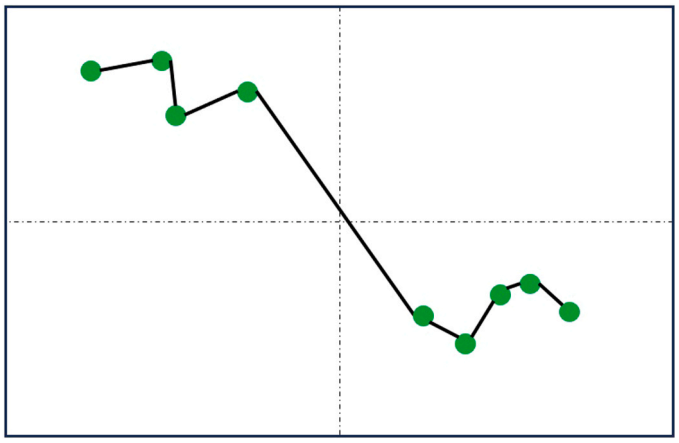

Figure B.11. Example of TS display locations during block v.

.

Appendix C. Simon Task

In the two-colour condition, the squares were coloured either red or blue. In the four-colour condition, the squares were coloured pink, green, yellow, or brown. Participants were instructed to respond the the stimuli as they appeared on the screen. For certain stimuli (red - two-colour condition, pink/green - four-colour condition), participants were required to use their left hand for responses, while for other stimuli (blue - two-colour condition, yellow/brown - four-colour condition), their right hand was to be used.

Throughout this work, a response to a single square is referred to as a trial. Trials were categorised into three groups based on the location of stimulus presentation: *congruent*, where the side of the displayed stimulus aligned with the side of the required response; *incongruent*, where the side of the stimulus presentation was opposite to the side of the intended response; and *neutral*, wherein the stimulus was centrally positioned on the screen, in keeping with Simon Task experiments carried out by [38,62].

A typical trial was as follows: the central fixation cross (FC) appeared (lasting 1500 ms), followed by the subsequent presentation of the target stimulus (TS) (displayed for 150 ms). As the TS disappeared, the central FC reappeared, denoting the initiation of the ensuing trial. For each trial, participants' responses were recorded in terms of accuracy and reaction time (RT). RT was recorded only for correct responses. In cases where participants did not respond within two seconds, the trial timed out and the response was classed as incorrect. Trials were presented in pseudorandom order to each participant.

The Simon task was comprised of three stages: **(i)** a practice test, twocolour condition, six trials, which participants could repeat as many times as they liked; **(ii)** two-colour condition, 24 trials (four repeats of each (*colour,location*) combination) **(iii)** four-colour condition, 48 trials (four repeats of each combination).

Pattern is a tuple of the form (*colour, location*). For the two-colour condition, possible colours are *b* and *r*, corresponding to blue and red, respectively; while possible colours in the four-colour condition are *y*, *p*, *g*, or *br* - yellow, pink, green, or brown, respectively. Correctness is a binary variable that defines whether or not the trial was correctly responded to - denoted by a 1 while a 0 indicates an incorrect trial. Latency details the RT for each correct button press. Note that the RT of incorrect trials, or time outs (no response after 2000 ms from TS presentation), were stored with a reaction time of 0.0 ms. Additionally, note that correct responses following an initial incorrect response were be marked as incorrect, but the latency of the correct response was recorded.

Appendix D. Exclusionary Criteria

Table D.11 details the full list of exclusionary criteria for the experiment, in the format in which it was shown to participants. Overarching categories are described, with further detail given in some cases. Note that wearing glasses, provided that this corrected eyesight to 20/20 vision, was not considered an exclusionary criterion.

Table D.11
Full list of experimental exclusionary criteria.

Category	Description
Vision Problems	Cataracts, diabetic retinopathy, macular degeneration, glaucoma, decreased visual field, poor night vision, poor colour vision.
Hearing loss	Right/left ear, both ears
Cardiovascular	
Stroke	
Transient Ischemic Attack (TIA)	
Fainting	
Neurological	Brain tumour, dementia, migraine or recurrent headaches, multiple sclerosis, Parkinson's Disease, Huntington's Disease, peripheral neuropathy, seizures, vertigo, dizziness, serious head injury (i.e. loss of memory or consciousness)
Psychiatric Illness	Depression, bipolar disorder, anxiety disorder, schizophrenia

(*continued on next page*)

Table D.11 (*continued*)

Category	Description
Attention deficient (hyperactivity) disorder	

Appendix E. Time Series Eye Movement Trace

Figure E.12 shows the eye movement trace of a single participant extracted from images captured during the experiment, with traces from V1 and V2 concatenated. The red lines indicate the portion of the video corresponding to the calibration routine.

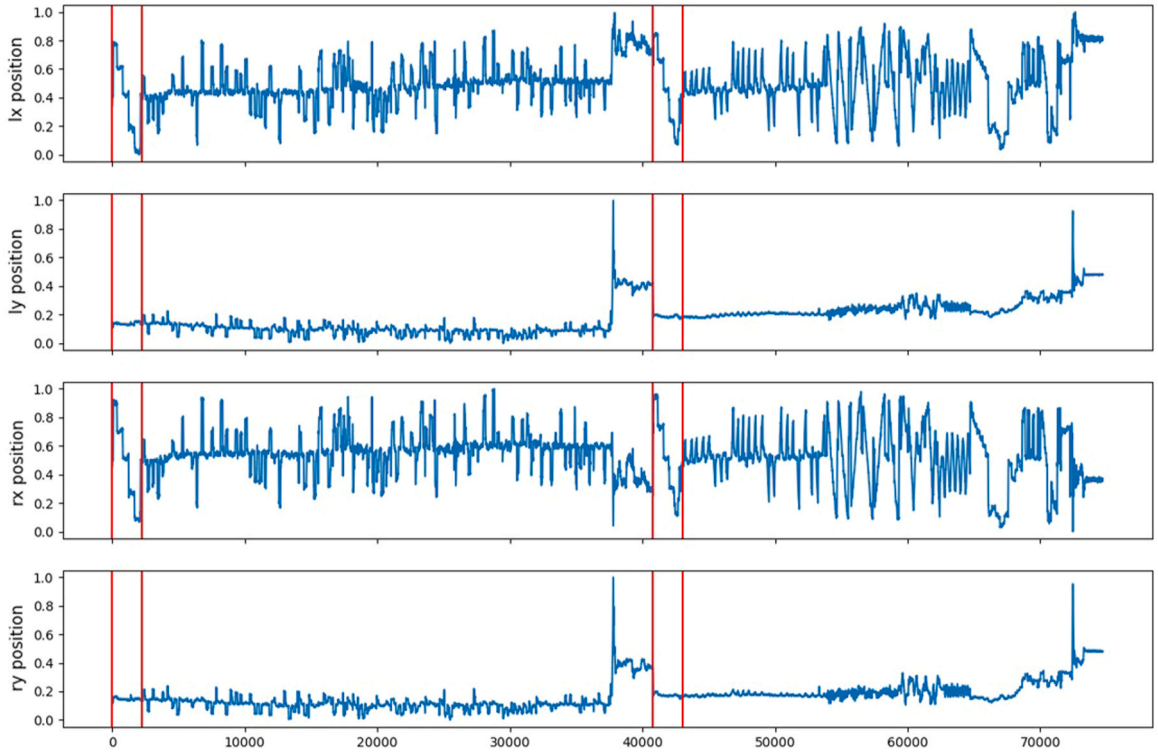

Figure E.12. Combined Saccadic and SPEM Eye Movement Time Series.

.

Appendix F. Time Series Model Architectures

The architectures for the FCN, Encoder, and Inception models used to classify cognitive load condition and age group are shown in Tables F.12, F.13, and F.14, respectively. The input to each model is assumed to be the saccadic (V1) trace. The architecture of the CNN model is shown in Table 5 and was comprised of two 1D convolution layers each followed by an average pooling layer, before the output from the final pooling layer was flattened.

The FCN was comprised of three consecutive 1D convolution layers followed by a global pooling and finally dense layer. Hyperparameters explored included the kernel size - (8,5,3) or (16, 10, 6) - along with the number of filters in each layer - (128, 256, 128) or (64, 128, 64).

For the Encoder and Inception models, hyperparameter optimisation was not carried out. As such, the tested models were exactly as illustrated in Tables F.13 and F.14, respectively.

The encoder model was composed of three convolutional blocks. The first two convolutional blocks consisted of a convolutional layer followed by max pooling, while the final block consisted of a convolutional layer followed by an attention layer [63]. The attention layer is fed to a fully connected layer, which is flattened before being fully connected to the output layer. Dropout of 0.2 applied after each convolutional layer, prior to max pooling.

Table F.12
FCN architecture. In this instance, the number of filters in the three consecutive convolutional layers are (128, 256, 128) respectively.

Layer	Output Shape
Input	(40800, 4)
1D Convolution	(40800, 128)

(*continued on next page*)

Table F.12 (*continued*)

Layer	Output Shape
1D Convolution	(40800, 256)
1D Convolution	(40800, 128)
1D Global Average Pooling	(128)
Fully Connected	1

Table F.13
Encoder architecture.

Layer	Output Shape
Input	(40800, 4)
1D Convolution	(40800, 128)
1D Max Pooling	(20400, 128)
1D Convolution	(20400, 256)
1D Max Pooling	(10200, 256)
1D Convolution	(10200, 512)
Attention	(10200, 256)
Fully Connected	(10200, 64)
Flatten	(652800)
Fully Connected	1

The InceptionTime model is constructed of two residual blocks, each containing three Inception modules, which are identical in overall structure to the Inception module depicted in [64].

The bottleneck layer of the Inception module has a kernel size of 1, while the following convolutional layers have kernel sizes of 1, 2, and 4, respectively. All convolutional layers contain 32 filters, and the max pooling layers have a pool size of 3.

Table F.14
Simplified Inception architecture.

Layer or Module	Output Shape
Inception Module	(40800, 128)
Inception Module	(40800, 128)
Inception Module	(40800, 128)
Inception Module	(40800, 128)
Inception Module	(40800, 128)
Inception Module	(40800, 128)
Global Average Pooling	(128)
Fully Connected	1

Appendix G. Hyperparameter Optimisation for CNN Models

The hyperparameters explored included: the number of convolutional blocks (defined as a convolutional layer followed by a pooling layer); the configuration of the ending layers (such as fully connected vs GAP layers); kernel sizes, and the number of filters in each layer.

Between two to four convolutional blocks were used, with performance higher for both tasks (distinguishing cognitive load condition, or between age groups) when using two and three block configurations. Ending layer configurations featured either: **(i)** a flatten layer, **(ii)** a global average pooling layer, **(iii)** fully-connected layers, following the final convolutional block. The output layer was always a fully-connected layer.

The kernel sizes and the number of filters explored for two, three, and four layer configurations are shown in Table G.15 and Table G.16, respectively.

Table G.15
The range of kernel sizes explored across time series models consisting of 2, 3, or 4 convolutional blocks.

Convolutional Blocks	Kernel Sizes
2	(7, 7)[1], (7, 9), (7, 13), (7, 17), (9, 13), (9, 17), (9, 19)
3	(3, 7, 9), (7, 7, 7), (7, 9, 13), (7, 9, 17), (9, 13, 17), (13, 17, 19)
4	(3, 7, 9, 13), (7, 9, 17, 19)

When distinguishing between cognitive load conditions, the ending layer configuration was most accurate when flattening the output of the convolutional layers, which fed into a single fully-connected output layer. Performance during cross validation was very poor when using either a global average pooling layer or fully-connected layers, with results typically achieving only 50–60% accuracy.

Table G.16
The different number of filters explored across time series models consisting of 2, 3, or 4 convolutional blocks.

Convolutional Blocks	Kernel Sizes
2	(2, 4), (2, 8), (6, 12), (8, 16), (8, 64), (16, 64), (64, 128)
3	(2, 8, 16), (2, 8, 64), (6, 12, 24), (8, 16, 32)
4	(6, 12, 24, 48)

When distinguishing between the youngest and oldest age groups, the ending layer configuration was most accurate when using a global average pooling layer prior to the output layer, exceeding performance when flattening the output of the convolutional layers. When multiple fully-connected layers were used, the model was prone to substantial overfitting. A three convolutional layer configuration was more accurate than two layer configurations, with four convolutional layers also resulting in substantial overfitting.

References

[1] Brookmeyer R, Evans DA, Hebert L, Langa KM, Heeringa SG, Plassman BL, Kukull WA. National estimates of the prevalence of alzheimer's disease in the united states. Alzheimer's Dement 2011;7(1):61–73.
[2] P. Moise, M. Schwarzinger, M. Um, Dementia care in 9 oecd countries: a comparative analysis oecd health working paper no. 13 (2004).
[3] C.G. Lyketsos, M.C. Carrillo, J.M. Ryan, A.S. Khachaturian, P. Trzepacz, J. Amatniek, J. Cedarbaum, R. Brashear, D.S. Miller, Neuropsychiatric symptoms in alzheimer's disease (2011).
[4] Porsteinsson A, Isaacson R, Knox S, Sabbagh M, Rubino I. Diagnosis of early alzheimer's disease: clinical practice in 2021, The journal of prevention of Alzheimer's. disease 2021;8:371–86.
[5] Gopalakrishna G, Brunton S, Pruzin J, Alford S, Hamersky C, Sabharwal A. Understanding the role of psychiatrists in the diagnosis and management of mild cognitive impairment and mild alzheimer's disease dementia: a cross-sectional survey. BMC Psychiatry 2023;23(1):716.
[6] Wolfe A. Institute of medicine report: crossing the quality chasm: a new health care system for the 21st century. Policy, Polit, Nurs Pract 2001;2(3):233–5.
[7] Bradford A, Kunik ME, Schulz P, Williams SP, Singh H. Missed and delayed diagnosis of dementia in primary care: prevalence and contributing factors. Alzheimer Dis Assoc Disord 2009;23(4):306.
[8] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR, Jr, Kaye J, Montine TJ, et al. Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimer's Dement 2011;7(3):280–92.
[9] Nichols E, Vos T. The estimation of the global prevalence of dementia from 1990-2019 and forecasted prevalence through 2050: an analysis for the global burden of disease (gbd) study 2019. Alzheimer's Dement 2021;17:e051496.
[10] Munoz D, Broughton J, Goldring J, Armstrong I. Age-related performance of human subjects on saccadic eye movement tasks. Exp brain Res 1998;121:391–400.
[11] Olincy A, Ross R, Youngd D, Freedman R. Age diminishes performance on an antisaccade eye movement task. Neurobiol Aging 1997;18(5):483–9.
[12] Schik G, Mohr S, Hofferberth B. Effect of aging on saccadic eye movements to visual and auditory targets. Int Tinnitus J 2000;6(2):154–9.
[13] Shafiq-Antonacci R, Maruff P, Whyte S, Tyler P, Dudgeon P, Currie J. The effects of age and mood on saccadic function in older individuals. J Gerontol Ser B: Psychol Sci Soc Sci 1999;54(6):P361–8.
[14] Port NL, Trimberger J, Hitzeman S, Redick B, Beckerman S. Micro and regular saccades across the lifespan during a visual search of "where's waldo" puzzles. Vis Res 2016;118:144–57.
[15] Morrow MJ, Sharpe JA. Smooth pursuit initiation in young and elderly subjects. Vis Res 1993;33(2):203–10.
[16] Sharpe JA, Sylvester TO. Effect of aging on horizontal smooth pursuit. Invest Ophthalmol Vis Sci 1978;17(5):465–8.
[17] Schewe H, Uebelhack R, Vohs K. Abnormality in saccadic eye movement in dementia. Eur Psychiatry 1999;14(1):52–3.
[18] Shafiq-Antonacci R, Maruff P, Masters C, Currie J. Spectrum of saccade system function in alzheimer disease. Arch Neurol 2003;60(9):1272–8.
[19] Crawford TJ, Higham S, Renvoize T, Patel J, Dale M, Suriya A, Tetley S. Inhibitory control of saccadic eye movements and cognitive impairment in alzheimer's disease. Biol Psychiatry 2005;57(9):1052–60.
[20] Kuskowski MA, Malone SM, Mortimer JA, Dysken MW. Smooth pursuit eye movements in dementia of the alzheimer type. Alzheimer Dis Assoc Disord 1989;3(3):157–71.
[21] Fletcher WA, Sharpe JA. Saccadic eye movement dysfunction in alzheimer's disease. Ann Neurol: J Am Neurol Assoc Child Neurol Soc 1986;20(4):464–71.
[22] Readman MR, Polden M, Gibbs MC, Wareing L, Crawford TJ. The potential of naturalistic eye movement tasks in the diagnosis of alzheimer's disease: a review. Brain Sci 2021;11(11):1503.
[23] Przybyszewski AW, Sledzianowski A, Chudzik A, Szlufik S, Koziorowski D. Machine learning and eye movements give insights into neurodegenerative disease mechanisms. Sensors 2023;23(4):2145.
[24] Oyama A, Takeda S, Ito Y, Nakajima T, Takami Y, Takeya Y, Yamamoto K, Sugimoto K, Shimizu H, Shimamura M, et al. Novel method for rapid assessment of cognitive impairment using highperformance eye-tracking technology. Sci Rep 2019;9(1):12932.
[25] Tadokoro K, Yamashita T, Fukui Y, Nomura E, Ohta Y, Ueno S, Nishina S, Tsunoda K, Wakutani Y, Takao Y, et al. Early detection of cognitive decline in mild cognitive impairment and alzheimer's disease with a novel eye tracking test. J Neurol Sci 2021;427:117529.
[26] Luke SG, Darowski ES, Gale SD. Predicting eye-movement characteristics across multiple tasks from working memory and executive control. Mem Cogn 2018;46:826–39.
[27] Broadbent DP, D'Innocenzo G, Ellmers TJ, Parsler J, Szameitat AJ, Bishop DT. Cognitive load, working memory capacity and driving performance: A preliminary fnirs and eye tracking study. Transp Res Part F: Traffic Psychol Behav 2023;92:121–32.
[28] Ross V, Jongen EM, Wang W, Brijs T, Brijs K, Ruiter RA, Wets G. Investigating the influence of working memory capacity when driving behavior is combined with cognitive load: an lct study of young novice drivers. Accid Anal Prev 2014;62:377–87.
[29] Stuyven E, Van der Goten K, Vandierendonck A, Claeys K, Crevits L. The effect of cognitive load on saccadic eye movements. Acta Psychol 2000;104(1):69–85.
[30] Chiarello M, Lee J, Salinas MM, Hilsabeck RC, Lewis-Peacock J, Sulzer J. The effect of biomechanical features on classification of dualtask gait. IEEE Sens J 2022;23(3):3079–89.
[31] Barhon LI, Batchelor J, Meares S, Chekaluk E, Shores EA. A comparison of the degree of effort involved in the tomm and the acs word choice test using a dual-task paradigm. Appl Neuropsychol: Adult 2015;22(2):114–23.
[32] Baddeley AD, Bressi S, Della Sala S, Logie R, Spinnler H. The decline of working memory in alzheimer's disease: a longitudinal study. Brain 1991;114(6):2521–42.
[33] Ai-brainscience, https://www.ai-brainscience.co.jp/en/, accessed: 2024–03-12.
[34] Viewmind, https://www.viewmind.com, accessed: 2024–03-12.
[35] Braingaze, https://braingaze.com, accessed: 2024–03-12.
[36] Leighton A, Weinborn M, Maybery M. Bridging the gap between neurocognitive processing theory and performance validity assessment among the cognitively impaired: a review and methodological approach. J Int Neuropsychol Soc 2014;20(9):873–86.
[37] Craft JL, Simon JR. Processing symbolic information from a visual display: interference from an irrelevant directional cue. J Exp Psychol 1970;83(3p1):415.
[38] Bialystok E, Craik FI, Klein R, Viswanathan M. Bilingualism, aging, and cognitive control: evidence from the simon task. Psychol Aging 2004;19(2):290.
[39] Van der Lubbe RH, Verleger R. Aging and the simon task. Psychophysiology 2002;39(1):100–10.
[40] Schmiedt-Fehr C, Schwendemann G, Herrmann M, Basar-Eroglu C. Parkinson's disease and age-related alterations in brain oscillations during a simon task. Neuroreport 2007;18(3):277–81.
[41] De Bruin A, Sala SD. Effects of age on inhibitory control are affected by task-specific features. Q J Exp Psychol 2018;71(5):1219–33.
[42] Lee S, Kawachi I, Berkman LF, Grodstein F. Education, other socioeconomic indicators, and cognitive function. Am J Epidemiol 2003;157(8):712–20.
[43] Lövdén M, Fratiglioni L, Glymour MM, Lindenberger U, Tucker-Drob EM. Education and cognitive functioning across the life span. Psychol Sci Public Interest 2020;21(1):6–41.
[44] Masley SC, Roetzheim R, Clayton G, Presby A, Sundberg K, Masley LV. Lifestyle markers predict cognitive function. J Am Coll Nutr 2017;36(8):617–23.
[45] Wu Y-T, Teale J, Matthews FE, Brayne C, Woods B, Clare L. Lifestyle factors, cognitive reserve, and cognitive function: results from the cognitive function and ageing study wales, a population-based cohort. Lancet 2016;388:S114.
[46] Zargari Marandi R, Madeleine P, Omland Ø, Vuillerme N, Samani A. Eye movement characteristics reflected fatigue development in both young and elderly individuals. Sci Rep 2018;8(1):13148.
[47] Miles G., Zhang W., Smith M. L., Zook N., EM-COGLOAD: Eye movement under varying cognitive loads across a range of age groups, doi: 10.17605/OSF.IO/ZJTDQ, https://osf.io/zjtdq/.
[48] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.

[49] Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput 2017;29(9):2352–449.
[50] Zhang W, Smith ML. Eye centre localisation with convolutional neural networks in high-and low-resolution images. in: International Conference on Computational Science and Its Applications. Springer; 2022. p. 373–84.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[52] F. Chollet, et al., Keras, https://keras.io (2015).
[53] Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mané D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viégas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X. TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org 2015. https://www.tensorflow.org/.
[54] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller P-A. Deep learning for time series classification: a review. Data Min Knowl Discov 2019;33(4):917–63.
[55] Zhao B, Lu H, Chen S, Liu J, Wu D. Convolutional neural networks for time series classification. J Syst Eng Electron 2017;28(1):162–9.
[56] O. Jesorsky, K.J. Kirchberg, R.W. Frischholz, Robust face detection using the hausdorff distance, in: Audio-and Video-Based Biometric Person Authentication: Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001 Proceedings 3, Springer, 2001, pp. 90–95.
[57] Turner S, Iliffe S, Downs M, Wilcock J, Bryans M, Levin E, Keady J, O'Carroll R. General practitioners' knowledge, confidence and attitudes in the diagnosis and management of dementia. Age Ageing 2004;33(5):461–7.
[58] Morgan DG, Crossley M, Kirk A, D'Arcy C, Stewart N, Biem J, Forbes D, Harder S, Basran J, Dal Bello-Haas V, et al. Improving access to dementia care: development and evaluation of a rural and remote memory clinic. Aging Ment Health 2009;13 (1):17–30.
[59] Mengoudi K, Ravi D, Yong KX, Primativo S, Pavisic IM, Brotherhood E, Lu K, Schott JM, Crutch SJ, Alexander DC. Augmenting dementia cognitive assessment with instruction-less eye-tracking tests. IEEE J Biomed Health Inform 2020;24(11): 3066–75.
[60] Robinson DA. The mechanics of human smooth pursuit eye movement. J Physiol 1965;180(3):569.
[61] Rashbass C. The relationship between saccadic and smooth tracking eye movements. J Physiol 1961;159(2):326.
[62] Aisenberg D, Henik A. Stop being neutral: simon takes control! Q J Exp Psychol 2012;65(2):295–304.
[63] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).
[64] Ismail Fawaz H, Lucas B, Forestier G, Pelletier C, Schmidt DF, Weber J, Webb GI, Idoumghar L, Muller P-A, Petitjean F. Inceptiontime: finding alexnet for time series classification. Data Min Knowl Discov 2020;34(6):1936–62.