Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Towards the automatic detection of activities of daily living using eye-movement and accelerometer data with neural networks

Jacob L. Newman [a],*, Zak Brook [a], Stephen J. Cox [a], John S. Phillips [b]

[a] School of Computing Sciences, University of East Anglia, University Drive, Norwich, NR4 7TJ, Norfolk, England, United Kingdom
[b] Otolaryngology, Norfolk and Norwich University Hospitals NHS Foundation Trust, Colney Lane, Norwich, NR4 7UY, Norfolk, England, United Kingdom

ARTICLE INFO

Keywords:
Activity classification
Time-series classification
Body-worn sensors
Digital biomarker

ABSTRACT

Early diagnosis of neurodegenerative diseases, such as Alzheimer's disease, improves treatment and care outcomes for patients. Early signs of cognitive decline can be detected using functional scales, which are written records completed by a clinician or carer, detailing a patient's capability to perform routine activities of daily living. For example, tasks requiring planning, such as meal preparation, are some of the earliest affected by early mild cognitive impairment. In this article, we describe work towards the development of a system to automatically discriminate and objectively quantify activities of daily living. We train a selection of neural networks to discriminate a novel list of 14 activities, specially selected to overlap with those measured by existing functional scales. Our dataset consists of eight hours of development data captured from four individuals wearing the Continuous Ambulatory Vestibular Assessment (CAVA) device, which was originally developed to aid the diagnosis of vertigo. Using frequency domain recognition features derived from eye-movement and accelerometer data, we compare several classification approaches, including three bespoke neural networks, and two established network architectures commonly applied to time-series classification problems. In 10-fold cross-validation experiments, a peak mean accuracy of 64.1% is obtained. The highest accuracy across all folds is 75.3%, produced by networks comprising Gated Recurrent Units. The addition of eye-movement data is shown to improve discrimination compared to using accelerometer data alone, by close to 9%. Classification accuracy is shown to degrade if the system is trained such that test subjects are excluded from the training data, with the small size of the dataset given as a likely explanation. Our findings demonstrate that the addition of eye-movement data can significantly improve the discrimination of daily activities, and that neural networks are well suited to this task.

1. Introduction

The United Kingdom has an ageing population, with the proportion of retirement age individuals expected to rise over the coming decades [1]. This demographic shift is expected to impact health and care services worldwide, as older individuals are increasingly living with multimorbidities, often impacting their ability to live independently [2,3]. For example, there is a concerning rise in the prevalence of degenerative brain diseases, such as Alzheimer's and Parkinson's disease [4]. These debilitating diseases force individuals to become entirely dependent on the care of others [5]. As with many diseases, early diagnosis is a key factor in their mitigation, by slowing disease progression through early access to treatment [6].

One of the earliest signs that an individual may be experiencing cognitive decline is a reduced capability to undertake routine daily tasks in their own home [7]. For example, individuals may show increased difficulty in undertaking tasks involving planning, such as meal preparation, organising finances or making travel arrangements [8]. They may also have difficulty in more routine tasks such as continence, personal hygiene and ambulating. Currently, the assessment of an individual's ability to perform activities of daily living (ADLs) is performed by a caregiver, clinician or even by the patient themselves [9,10]. The focus of the work in this article is to work towards the development of a system for automatically quantifying ADLs, with a view to providing a proxy for monitoring cognitive decline. The current, manual process typically requires the completion of a form designed to measure the degree of function with respect to routine daily activities, commonly referred to as functional scales [11]. There are broadly two scales used for this purpose: The Katz Index [12], which covers a range of basic activities, and the Lawton scale of Instrumental Activities of Daily Living (IADL) [13], which includes more complex tasks, such as those requiring planning.

* Corresponding author.
E-mail address: jacob.newman@uea.ac.uk (J.L. Newman).

https://doi.org/10.1016/j.compbiomed.2024.109607
Received 23 September 2024; Received in revised form 19 December 2024; Accepted 19 December 2024
Available online 1 January 2025

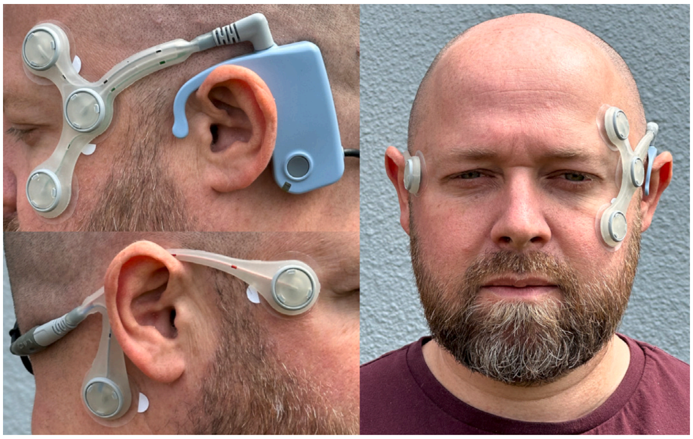


Fig. 1. The CAVA device. Electrodes adhered to the face capture the corneo-retinal potential generated by the eyes, with data recorded on the logging unit behind the left ear. The device also records acceleration of the head using an accelerometer in the logging unit, an LED provides an indication of device status, and a button on the device can be used to mark times of interest in the data recorded.

Although widely used, manual records of ADLs and IADLs are subjective and inaccurate, as they are typically completed retrospectively and based on the judgement of one individual [14]. There is no universally agreed scale on which to assess ADLs, and the variety of scales that do exist mean that patients are not assessed in a consistent manner [15]. We work towards a solution to these challenges by exploring the feasibility of using body-worn sensors and machine learning to obtain automatic and objective records of ADLs, potentially enabling more accurate and informative assessments of an individual's cognitive decline and ability to live independently. Effectively, the data generated by this process could be used as a digital biomarker, enabling the objective measurement of the effectiveness of interventions for neurodegenerative diseases, such as drug therapies, physiotherapy and social support. In a similar vein, it could also be used to assess clinical trial outcomes for new treatments. Although our intended application is primarily the measurement of cognitive decline, this work also has potential application to the detection of abnormal events occurring within an individual's daily routine, such as tonic-clonic epileptic seizures [16] or syncope [17], or for assessing ADLs in stroke-affected patients [18].

This work is undertaken using data captured with the Continuous Ambulatory Vestibular Assessment (CAVA) device. The CAVA device was originally developed to aid the diagnosis of a specific type of dizziness called vertigo [19,20]. The device is worn on the face and records horizontal and vertical eye-movements, as well as three-axis acceleration of the head (Fig. 1). Patients suspected of suffering from vertigo would wear the device nearly continuously for up to a month. After the monitoring period, the data from the device would be analysed by computer algorithms for evidence of signals indicating vertigo. Using this device, in this work we examine whether the physiological parameters it records could aid the automatic and objective detection of ADLs. The ability to classify and quantify a timeline of daily activities would effectively provide a digital biomarker, which could be used to monitor changes in an individual's behaviour in response to treatment. CAVA is particularly suited to this given that it can be worn throughout an individual's normal daily life: nearly continuously, throughout the day and night, and for up to a month. Also, unlike other devices which provide accelerometer data alone, the CAVA device records eye-movements, which may aid in the detection of more complex activities that do not require physical movement of the body. While this is not the first work to explore the classification of ADLs using both eye- and head-movement features [21], as far as the authors are aware, it is the first to apply contemporary neural network approaches to this type of the data for this specific task.

Classification of human activities from sensor data is a mature field of research. For detailed reviews of studies in this area, please refer to [22,23]. Approaches using body-worn sensors to discriminate activities either make use of modern smartphone technology, which typically includes auditory and accelerometer-based sensors, or uses separate wearable sensors, recording parameters such as heart rate, blood oxygen saturation and gyroscopic or accelerative movement. The majority of previous work has involved the use of accelerometer data alone. An objective of our work is to determine if the inclusion of eye-movement data improves the capabilities of an automatic system, motivating the use of the CAVA device for this application.

A variety of traditional machine learning approaches have been applied to activity classification, including support vector machines (SVMs), K-nearest neighbour (KNN), decision trees and neural networks. In [24], six basic activities such as walking, sitting, and ascending stairs were considered, with data captured from nine individuals. They explored a selection of classifiers such as SVMs and KNN, obtaining classification accuracies in the range 53.1% to 99.4%. A similar study by [25], achieved a peak mean accuracy of 93.38% using SVMs. In a three activity classification task involving sitting, walking and standing, [26] achieved accuracies in the range of 76.2% to 97.9% using random forest, naive Bayes and KNN classifiers. In [27], seven different activities, including lying, sitting, walking, Nordic walking, running, rowing, and cycling were considered. They collected 31 h of labelled data, measuring 18 different parameters, such as heart rate, acceleration of the wrist, and location through GPS. Eye-movement was not recorded. Decision trees and neural networks were evaluated and a maximum mean accuracy of 86% was obtained. There is debate in the community regarding the appropriateness of subject-independent classifiers, for which the test subjects are not included in the training data. Because of the variability of physiological signals, it has been suggested that person-specific models are required for optimal results [28].

These previous studies demonstrate the feasibility of using body-worn sensors to classify activities and show that a high degree of accuracy can often be obtained for distinctive, simple activities. They also highlight the variability between data captured from different individuals, leading to poor subject-independent classification performance. The number of activities considered in previous studies is often small, and those that are selected are very distinctive, providing little challenge for contemporary machine learning methods. Hence, classification accuracies reported are often in excess of 95%. To avoid this pitfall, we intend to extend the work of these previous studies by performing a classification task involving a larger selection of 14 activities, designed to overlap with the more complex and nuanced activities defined by the ADL and IADL definitions. We will evaluate the capability of our algorithms to detect this novel list of activities, in the hope that others in the field may also explore its suitability for this problem. Our own previous work involving eye-movement data, and specifically data from the CAVA device, has shown that neural networks are well suited to this type of problem, in that they are able to learn the complex, temporal relationships that exist within physiological signals that more traditional methods may fail to. We will therefore present the results of applying neural networks to the challenge of activity classification.

The remainder of this article is organised as follows: In Section 2 we introduce the dataset used in our experiments. Sections 3 and 4 present details of the system developed, in particular, the feature extraction method and neural network architectures used in this work. In Section 5 we describe the experimental framework used to evaluate the system developed, including details of the neural network architecture used. Results are provided in Section 6, with Discussion and Conclusions presented in Sections 7 and 8.

2. Dataset

A small quantity of development data was collected for use in this work. This data was captured using the CAVA device, which was worn by four study members directly involved with this project, who each wore the device for less than a day in total. The study members were all male and aged between 20 and 40. As described, the CAVA device was originally developed as a diagnostic tool for detecting abnormal eye-movements in patients suspected of suffering from vertigo. It samples horizontal and vertical eye-movements at ~42 Hz as well as three-axis accelerometer data at 8 Hz. A button on the device can be used by patients and clinicians to mark the time of events of interest, or simply to provide a useful reference point (e.g. The time at which the device was donned). An extended button press can be used to check for the correct functioning of the device, together with a status indicator LED.

As we work towards developing a complete system for automatically classifying ADLs, we opted to focus on activities that align closely to those defined by the existing functional scales. These ADLs fall broadly into two categories: basic ADLs [12] and instrumental ADLs (IADLs) [13]. A list of the activities included in these definitions is shown in Table 1, but to summarise, basic ADLs include activities such as ambulating, feeding, and personal hygiene, while IADLs include more complex activities, often requiring more planning by the individual. The capability to perform IADLs is more often impacted by cognitive impairment [29]. Using these definitions, we defined our own set of 14 activities, which we term CAVA ADLs (CADLs) (See Table 2). These were selected to be challenging to discriminate using body-worn sensors. In cases where identification of a specific activity was likely to be impossible using the available sensor data, such as for managing finances, we opted to use a simpler but related activity, or to exclude the activity from our definition. Table 1 shows how the CADLs relate to conventional ADLs and IADLs.

During the data capture process, project members recorded the data without assistance and in their own home. They were given only simple written and verbal instructions on how to undertake each activity, in order that the activities might reflect some real-world variability in how they were performed. Table 3 shows the quantity of data recorded by each individual for each activity. Fig. 2 shows examples of the eye-movement and accelerometer data corresponding to *reading*, *eating* and *ambulating*.

Training neural networks to detect the target activities required the data to be labelled at a sample level. To aid this process, each participant was instructed to keep a record of the activities they undertook. They used the CAVA device's event marker to signify the start and of each activity, and they kept a written record of the corresponding times and type of activity. The data labelling was performed manually by one individual. This process involved plotting the data for each participant, manually identifying the event markers corresponding to the written records, and storing the sample-level indices for each activity's start and end point. A simple MATLAB script was created to iterate through the stored indices and to generate a corresponding label vector. The final vector was of equal length to the sampled eye-movement data, with each element containing either the name of the activity present or a '0', indicating no activity at that sample.

Table 1
Showing the relationship between Activities of Daily Living (ADL), Instrumental ADL (IADL) and CAVA ADL (CADL) definitions.

Activity	ADL	IADL	CADL	CADL activity
Ambulating	[a]		[a]	Ambulating
Dressing	[a]		[a]	Dressing
Personal hygiene	[a]		[a]	Washing hands
			[a]	Brushing teeth
Feeding	[a]		[a]	Eating
Toileting/Continence	[a]			
Meal preparation		[a]	[a]	Chopping
			[a]	Frying
Housecleaning		[a]	[a]	Washing dishes
Finances		[a]	[a]	Mobile admin
Communication		[a]	[a]	Reading
			[a]	Writing
			[a]	Typing
Transportation		[a]		
Medication		[a]		
			[a]	TV
			[a]	Short-form videos

[a] Indicates that an activity is present in the definition.

Table 2
Definition of activities included in the CAVA ADL (CADL) dataset.

Activity	Definition
Ambulating	Walking around the home
Dressing	Putting clothes on and taking them off
Washing hands	Using soap and water to wash hands
Brushing teeth	Using a manual toothbrush
Eating	Eating a meal
Chopping	Chopping food in preparation for eating
Frying	Cooking food in a frying pan
Washing dishes	Washing dishes in a sink using soap and water
Mobile admin	General administration on a mobile phone, such as emails
Reading	Reading text from a book or paper
Writing	Handwriting
Typing	Typing text on a keyboard
TV	Watching videos on a screen or television
Short-form videos	Watching TikTok/Instagram videos on a mobile phone

Table 3
Quantity of data (in minutes) recorded by each individual for each of the CADL activities.

Activity	Subject 1	Subject 2	Subject 3	Subject 4	Total
Ambulating	25	6	1	0	32
Dressing	2	4	6	0	12
Washing hands	1	4	4	0	9
Brushing teeth	2	5	6	2	15
Eating	5	6	9	0	20
Chopping	9	5	6	0	20
Frying	13	6	6	0	25
Washing dishes	10	11	12	0	33
Mobile admin	9	5	12	21	47
Reading	31	10	9	20	70
Writing	5	11	4	0	20
Typing	6	15	7	43	71
TV	40	16	23	11	90
Short-form videos	9	11	3	0	23
Total	167	101	108	97	473

3. Feature extraction

The data from the CAVA device includes both two-channel eye-movement and three-channel accelerometer data. As the accelerometer data was sampled at a lower rate than the eye-movement data, the accelerometer data consisted of fewer samples. To enable the vectors for each data channel to be stacked into a feature matrix, we interpolated the accelerometer data using spline fitting, after which the number

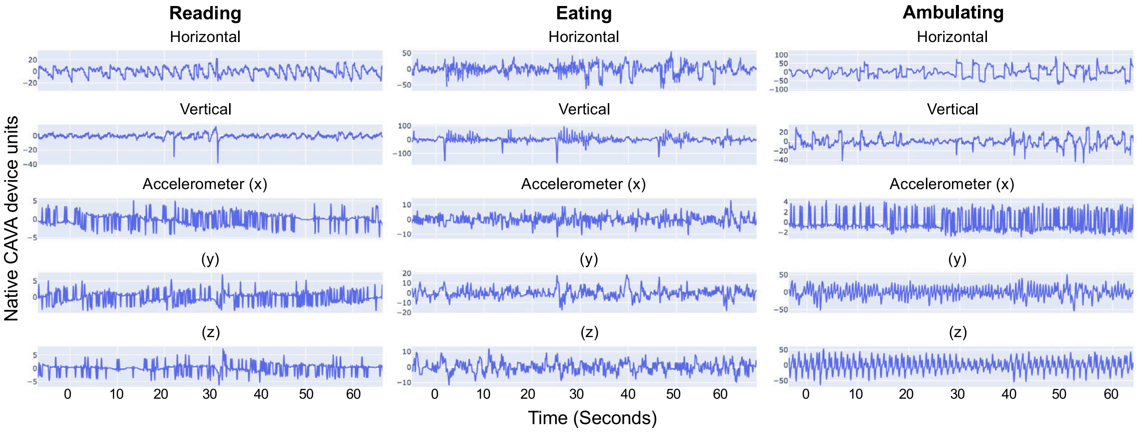


Fig. 2. Example horizontal and vertical eye-movement and accelerometer (x, y, z) data for *reading, eating* and *ambulating* activities. Reading is distinctive by the saw-tooth-like horizontal eye-movement signal, eating by a high frequency vertical eye-movement component, and *ambulating* by periodic accelerometer signals, particularly in the z channel.

of samples in each channel was equal. No further pre-processing was performed.

The data from all subjects were divided into non-overlapping segments of a fixed duration. We chose to use non-overlapping segments as this reduces the dependency between adjacent segments, since the same data cannot appear in two segments simultaneously. The majority of our experiments used a segment duration of three seconds (125 samples), which was determined through experimentation to be the optimal value (See Section 5). The class label attributed to each segment was determined by the mode class label of the samples within it.

In this work, we chose to use frequency domain recognition features, by separately applying the Fast Fourier Transform (FFT) to each vector within each feature matrix. Such features have been widely used in a variety of signal processing applications including speech recognition [30] and image processing [31], and have previously been applied to activity classification tasks [32]. Using this approach has two main advantages. Firstly, movement activities consist of complicated, repetitive motions, which should be distinctive in the frequency domain. Secondly, by using the magnitude spectrum, we will reduce the dimensionality of our data, which may be advantageous given the relatively small size of our dataset.

The magnitude spectrum was calculated for each sample, for each data channel, from the complex spectrum resulting from applying the FFT. We discarded the second half of the frequency bins in the magnitude spectrum, as they are a reflection of the first half and thus redundant. The frequencies retained were in the range of 0 Hz to ∼21 Hz. In our experiments using three second segments, this equates to a frequency resolution of 0.33 Hz per bin. After application to each segment, the resulting feature matrix consisted of five rows (one for each data channel) and 62 columns (each bin of the magnitude spectrum). Each row was normalised to a unit vector by dividing each element by the magnitude of the vector.

4. Neural networks

In this work, we evaluate three custom neural network architectures and also two established networks: Fully Convolutional Networks (FCNs) [33] and ResNet [34]. Our own networks were developed using Google's TensorFlow [35] and Python. For implementations of the FCN and ResNet architectures, we used the Python code provided by the authors of [36]. Full details of the neural network architectures we developed for this work are presented in Fig. 3, including default network hyperparameters and details of ranges explored during hyperparameter tuning.

As described in Section 5, most experiments were conducted as 10-fold cross-validation experiments. One experiment used hold-one-subject-out cross-validation, and a further simply divided the data into training, validation and testing partitions. Using the KerasTuner Python library [37] and the Hyperband tuning algorithm, extensive hyperparameter tuning was carried out for each fold in the cross-validation experiments, using the validation data to determine the optimal hyperparameters. In Fig. 3, the ranges of hyperparameters explored during tuning are provided on the left-hand side of the figure. The specific set of hyperparameters explored depended on the network being trained. For example, one architecture did not contain convolutional layers, and hence there was no requirement to tune the number of convolutional kernels. We did not tune the hyperparameters for the FCN and ResNet architectures, as these networks are already optimised. Also, for our first experiment exploring segment duration, we used the default hyperparameter values shown on the right-hand side of Fig. 3.

Our base network consists of three 1D convolutional layers, followed by a dropout layer to reduce overfitting. As our data is time-series data, for the later layers we explored architectures typically used for time-series classification problems. Namely, we compared the effectiveness of Long Short-Term Memory layers (LSTM) [38] and Gated Recurrent Units (GRU) [39]. These architectures are known to be able to capture the temporal dependencies in data, in a way which conventional, fully-connected layers cannot. Therefore, experiments using our own neural network architecture either make use of LSTM or GRU layers (illustrated by the pink and green shaded boxes in Fig. 3). When considering the GRU layers, we also examined whether the initial convolutional layers added to the network's discriminatory capabilities by performing experiments both with and without the initial convolutional layers (Indicated by the blue shaded box in Fig. 3). The output layers of the network consisted of two fully-connected layers.

Each of our own networks was trained to a maximum of 500 epochs. The Adam optimiser was used and the categorical cross-entropy loss metric. Early stopping criteria was used to halt training if the validation loss did not decrease for 50 successive epochs. The weights retained at the end of training were the ones producing the highest validation accuracy. All neural networks were trained using CPU and GPU resources on the University of East Anglia's High Performance Computing cluster.

As Table 1 shows, our dataset does not include an equal number of samples for each class of activity. A class imbalance such as this can impede the capabilities of a classifier to learn to discriminate the minority classes. The ideal solution is to balance the classes by recording more data for the minority classes. Instead, for each network trained, we used Synthetic Minority Oversampling Technique (SMOTE) to balance the quantity of samples within each class of the training set. SMOTE works by synthesising new data points for the minority classes by sampling the feature space between neighbouring data points [40]. Note we do not balance the test data as it would be misleading to report the accuracy obtained using synthesised data samples.

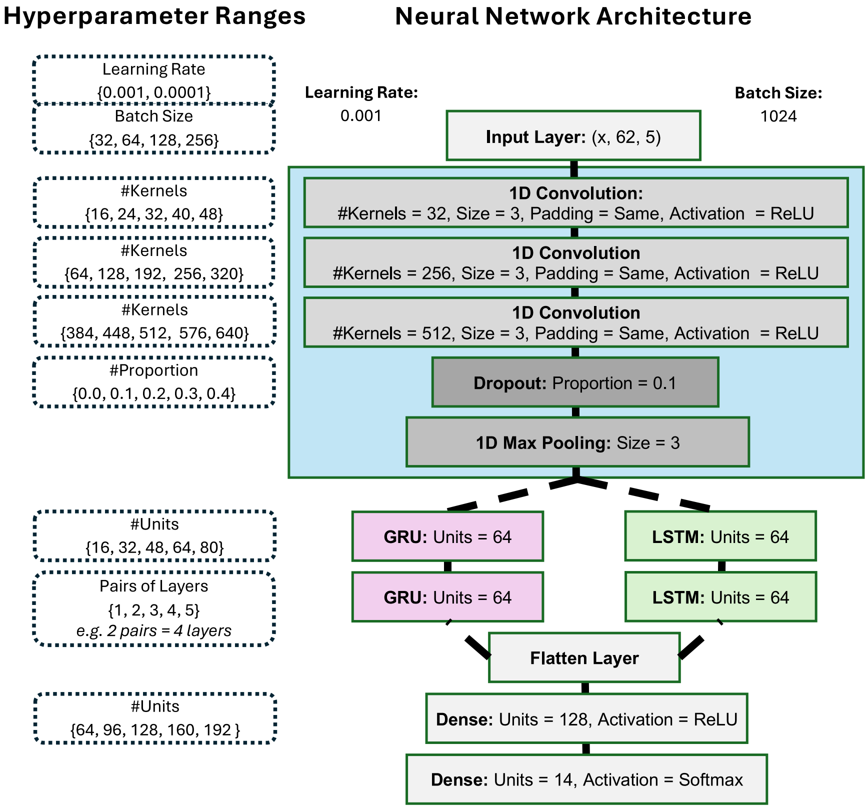


Fig. 3. The right-hand side of the figure shows the neural network architecture implemented in TensorFlow and used in our experiments. Default network hyperparameters are displayed. The thick, dashed lines indicate two types of architecture used in separate experiments: either a pair of GRU layers or a pair of LSTM layers. Experiments using the GRU layers are conducted both with and without the use of the earlier convolutional layers, which are shown within a blue box. The left-hand side of the figure shows the hyperparameter ranges explored during hyperparameter tuning.

5. Experiments

We designed a range of experiments to determine the suitability of the system developed. All experiments, except the first experiment to determine the optimal segment duration, were conducted as cross-validation experiments. 10-fold cross-validation was used for all experiments except for the subject-independent experiment, in which hold-one-subject-out cross-validation was used. Stratified sampling was used to generate the 10-fold partitions, and to create all validation data partitions. For each training fold in these experiments, 10% of the training data was used as validation data.

Experiments were undertaken in either a multi-subject or subject-independent training mode. In the multi-subject experiments, the data were randomly sampled from all available samples, and therefore, it was possible for data from each subject to appear in each of the training, validation and testing sets. This experiment was intended to explore the potential capabilities of our classification system, if trained using a larger, more diverse dataset. In the subject-independent experiments, test data was exclusively taken from the specific subject being tested, with training and validation data drawn from the remaining subjects. As before, 10% of the training data was used as validation data. This scenario evaluated how our system might perform if presented with data from new individuals, reflecting the potential real-world use of this system.

The first experiment aimed to determine the optimal segment duration for use in subsequent experiments. We trained networks using segment durations of between 1 and 10 s, corresponding to between 42 and 417 individual samples. Changing the segment duration also varied the total number segments available for training, validation and testing (See Table 4). The neural network used in this experiment contained GRU layers, as described in Section 4. This experiment was undertaken in a multi-subject mode. In this experiment, 10% of the total available data was held out for validation purposes, and 10% for testing, with the optimal duration determined from the results obtained for the validation data. Our intuition was that the optimal duration would be long enough to contain signals representative of each activity, but short enough to maximise the number of data samples available for training our networks.

Next, we performed further multi-subject classification experiments designed to compare two common approaches to time-series classification problems: GRU and LSTM neural network architectures. We also examined the performance of the GRU network without prior convolutional layers. Additionally, we evaluated two established neural network approaches to time-series classification: Fully Convolutional Networks (FCNs) and ResNet.

The final experiment sought to determine if eye-movement data, which is not typically used for activity classification, provides improved discrimination of activities compared to accelerometer data alone. This was achieved by training three separate neural networks; one using eye-movement data alone, one using accelerometer data alone, and a final network using a combination of both eye-movement and accelerometer data. The results obtained by these three networks are compared to determine the discriminative power of each data channel.

6. Results

First, we present the results of experiments in which the segment duration was varied. The plot in Fig. 4 shows that as the segment duration increased from one second to three seconds, the mean classification accuracy also increased from 70.0% to 76.1%. Beyond this value, with segment durations of between four and ten seconds, the classification accuracy fluctuated between 69.7% and 76.6%. Three seconds was selected as the optimal segment duration, as this value produced a high classification accuracy whilst also providing a greater number of

Table 4
Quantity of segments in each data partition for given segment durations.

Segment Duration (S)	# Training samples	# Validation samples	# Testing samples
1	20 701	4436	4436
2	10 225	2192	2191
3	6790	1455	1455
4	5112	1096	1096
5	4080	875	874
6	3395	728	727
7	2916	625	625
8	2548	547	546
9	2263	485	485
10	2039	438	437

Table 5
Classification accuracies (%) for 10-fold cross-validation experiments. The first five columns relate to experiments using the different neural network architectures described in Section 4. The last two columns relate to experiments using the GRU network architecture with different feature sets: either accelerometer features or eye-movement features alone.

Fold	LSTM	GRU w/o Conv.	GRU	FCN	ResNet	Accel. Feats. Only	Eye Feats. Only
1	38.87	41.24	42.68	43.61	46.70	47.42	31.55
2	60.10	57.63	63.40	62.27	69.69	58.45	44.85
3	66.08	61.34	67.11	56.91	64.02	52.16	48.87
4	70.52	64.64	61.34	60.93	69.59	68.25	46.29
5	58.04	59.07	57.01	49.79	56.70	53.09	47.84
6	67.42	60.10	66.39	56.70	68.04	54.23	49.38
7	74.33	66.91	75.26	65.05	71.96	61.34	57.22
8	68.56	65.36	74.33	58.76	68.56	57.84	64.33
9	64.43	60.93	65.05	51.34	63.40	51.96	52.58
10	63.20	48.66	67.94	48.04	61.96	47.63	53.71
μ (%)	63.15	58.59	64.05	55.34	64.06	55.24	49.66
σ (%)	9.79	7.94	9.29	6.54	7.22	6.40	8.59

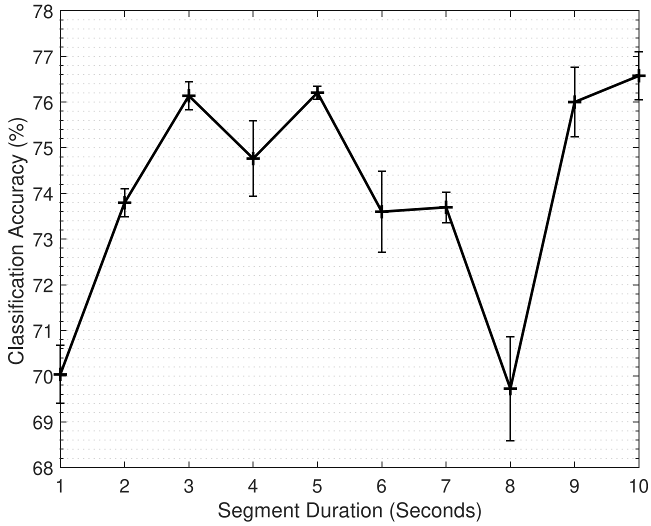


Fig. 4. Plot showing the mean accuracies obtained using different segment durations. Error bars indicate standard error of the mean for five repetitions of each experiment. The only differences between these repeats are the random weights and biases that each network is initialised with at the start of training.

segments for training, validation and testing purposes (Compare the number of three second segments with ten second segments in Table 4).

Results for subsequent experiments are shown in Table 5 and Fig. 5. To determine the statistical significance of each result, we first conducted a Kolmogorov–Smirnov test for normality on each set of ten results. For every set of experiments, we accepted the null hypothesis that the results were normally distributed, at $\alpha = 0.05$. Following this, for each pair of result sets we compared, we conducted a one-tailed, paired samples t-test, using $\alpha = 0.05$.

The mean classification accuracy obtained using the network including LSTM layers was 63.15%. The network using GRU layers achieved a higher accuracy of 64.05%, although this difference was not statistically significant. This set of experiments yielded the highest individual accuracy for any fold tested, of 75.26% for fold #7. When the earlier convolutional layers were removed from the GRU network, the mean accuracy decreased to 58.59%, and this difference was statistically significant. We also evaluated two established network architectures, FCN and ResNet. FCN was the worst performing network when using both accelerometer and eye-movement features, achieving a mean accuracy of 55.34% across all ten folds. ResNet gave a mean accuracy of 64.06%, which the highest mean accuracy achieved across all experiments, and which was marginally higher than that obtained using our own network containing GRU layers (64.06% compared to 64.05%). This difference was not statistically significant. Thus, the GRU and ResNet networks were the highest-performing in these experiments. Given the similarity in classification performance between them and that the GRU network out-performed ResNet for 6 out of 10 folds, we opted to use the GRU architecture in all subsequent experiments.

The confusion matrix shown in Fig. 6 relates to the experimental repeat producing the highest accuracy, fold #7 using the GRU architecture. The matrix shows a high level of classification accuracy is obtained for many of the activity classes, as indicated by the high values along the diagonal of the matrix. For example, activities such as *writing*, *ambulating*, and *eating* are discriminated with high accuracy. However, some patterns of confusion are observed, such as for *reading* and watching of *short-form videos*, which are confused with *mobile admin*. Also *typing*, which is confused with *reading* and watching *TV*, and *washing dishes*, which is confused with *chopping*, *washing hands*, and *reading*. We will discuss these results in Section 7.

In Table 6 and Fig. 7, we present the results of our subject-independent experiments. The results reveal a significant drop in classification accuracy when the test subject is not included in the training set. The best performing classifier was for subject number three, which achieved a mean accuracy of 30.43%, while the lowest mean accuracy was 17.71%, obtained for subject number four. The results for each of the five repeat experiments are shown in Fig. 7, and these reveal a narrow spread of values, indicating fairly consistent performance between training runs.

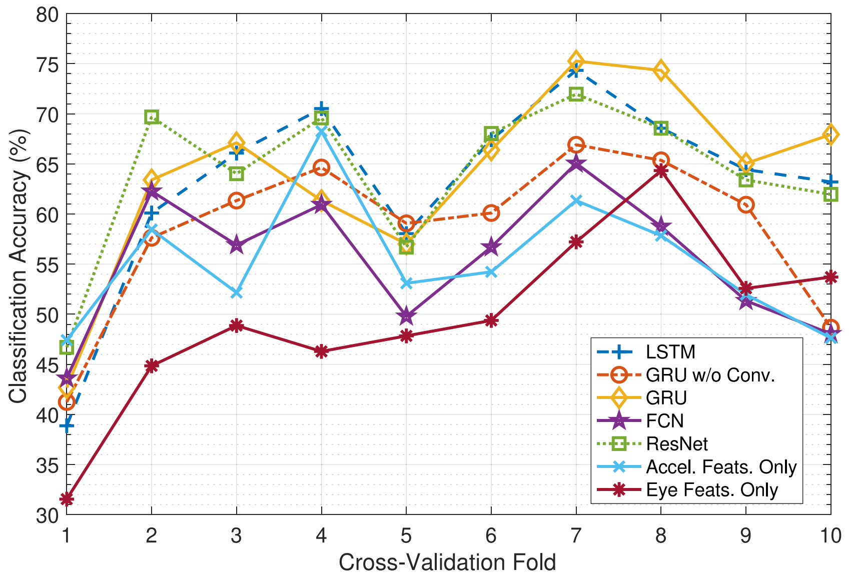
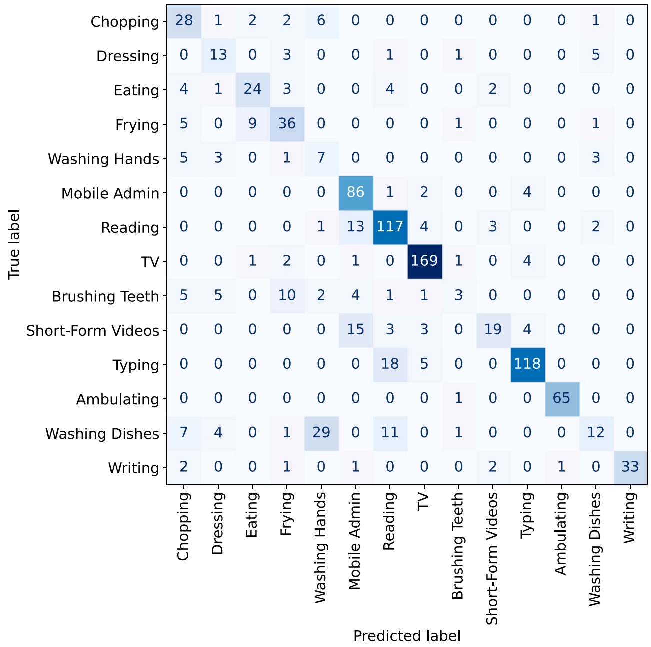


Fig. 5. Showing the individual classification accuracies (%) for each network architecture and feature sets considered, for each of the 10 cross-validation folds.



	Chopping	Dressing	Eating	Frying	Washing Hands	Mobile Admin	Reading	TV	Brushing Teeth	Short-Form Videos	Typing	Ambulating	Washing Dishes	Writing
Chopping	28	1	2	2	6	0	0	0	0	0	0	0	1	0
Dressing	0	13	0	3	0	0	1	0	1	0	0	0	5	0
Eating	4	1	24	3	0	0	4	0	0	2	0	0	0	0
Frying	5	0	9	36	0	0	0	0	1	0	0	0	1	0
Washing Hands	5	3	0	1	7	0	0	0	0	0	0	0	3	0
Mobile Admin	0	0	0	0	0	86	1	2	0	0	4	0	0	0
Reading	0	0	0	0	1	13	117	4	0	3	0	0	2	0
TV	0	0	1	2	0	1	0	169	1	0	4	0	0	0
Brushing Teeth	5	5	0	10	2	4	1	1	3	0	0	0	0	0
Short-Form Videos	0	0	0	0	0	15	3	3	0	19	4	0	0	0
Typing	0	0	0	0	0	0	18	5	0	0	118	0	0	0
Ambulating	0	0	0	0	0	0	0	0	1	0	0	65	0	0
Washing Dishes	7	4	0	1	29	0	11	0	1	0	0	0	12	0
Writing	2	0	0	1	0	1	0	0	0	2	0	1	0	33

Fig. 6. Confusion matrix for the classification of test samples for fold #7 of experiments using GRU layers (See Table 5 and Fig. 5).

Table 6
Classification accuracies (%) for subject independent experiments, in which a separate network is trained for each of the four test subjects using data from the remaining three subjects. These experiments were repeated five times, with the only differences being the random weights and biases that each network were initialised with at the start of training.

Subject ID	1	2	3	4
Repeat 1	25.45	19.79	28.35	18.08
Repeat 2	26.29	24.56	33.12	16.68
Repeat 3	24.65	22.82	31.54	18.03
Repeat 4	26.26	25.68	30.71	17.77
Repeat 5	22.36	26.49	28.45	17.98
μ (%)	25.00	23.87	30.43	17.71
σ (%)	1.62	2.66	2.05	0.58

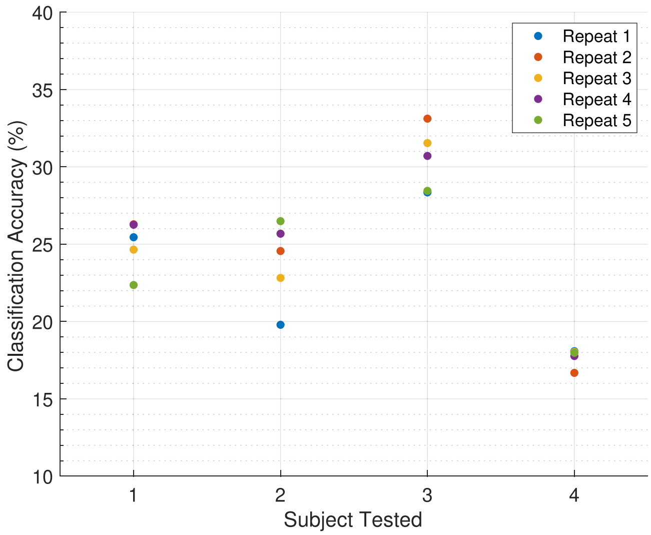


Fig. 7. A plot showing the classification accuracy for subject-independent experiments, in which a separate neural network was trained for each test subject using only data from the other three subjects.

The last two columns in Table 5 relate to experiments involving neural networks trained using recognition features derived from accelerometer data alone or from eye-movement data alone. The use of eye-movement data alone produced a mean classification accuracy of 49.66%, while accelerometer data alone gave a mean accuracy of 55.24%. Notably, both of these results are poorer than that obtained using a combination of these data sources, at 64.05%.

Lastly, the confusion matrix in Fig. 8 shows the classification differences between the repeat experiment using accelerometer data alone compared to the best performing network using a combination of both data channels (As shown in Fig. 6). The positive values across the diagonal indicate an increase in correct classifications for a number of activities, while the negative values in the off diagonals reveal a decrease in confusions. These results show that the addition of eye-movement data improved the network's discrimination of activities including *eating*, *mobile admin*, *reading*, *TV*, watching *short-form videos* and *typing*, i.e. mostly activities in which eye-movement is the dominant

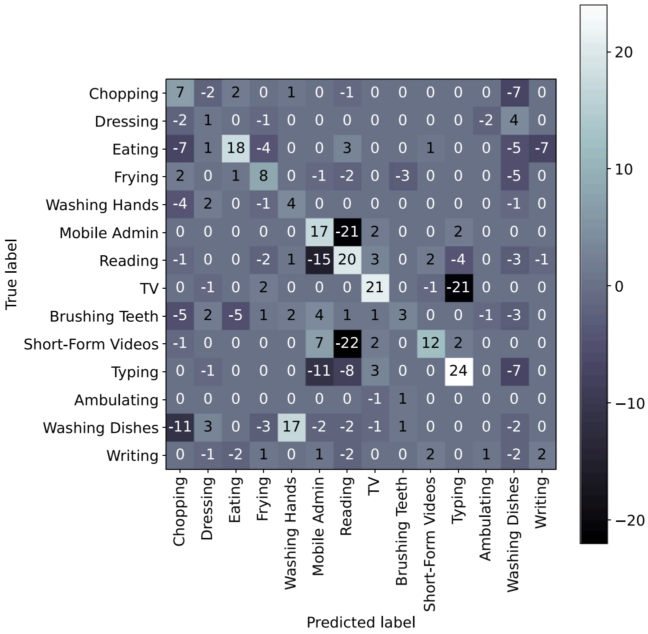


Fig. 8. Confusion matrix showing the difference between the best performing network trained using combined eye-movement and accelerometer derived recognition features and the same network architecture trained using accelerometer data alone. Positive values indicate an increase for a given confusion or, along the diagonal, for the number of correct classifications. Negative values indicate a decrease in classifications. Values of zero indicate no change.

movement. The addition of eye-movement did lead to an increase in confusions between *washing dishes* and *washing hands*.

7. Discussion

After the optimal segment duration had been confirmed in Section 6, the first set of results compared the accuracies obtained by two different but commonly used architectures for time-series classification tasks: LSTM and GRU layers. The results confirmed that both approaches could be used to discriminate the 14 activities considered here and with a good degree of accuracy. Confirming results from other unrelated studies, the networks using GRU layers outperformed the more commonly used LSTM approaches, although the difference was not statistically significant [41,42]. This result also accords with the authors' own experiences with other time-series data classification problems, and is perhaps explained by the reduced complexity of GRU layers being better suited to the relatively small dataset used in this work. The inclusion of earlier 1D convolutional layers is shown to contribute significantly to the discriminative power of these networks. Our own network configurations are shown to give comparable results to those obtained using ResNet, and better results than using FCNs.

Some of the confusions observed for the repeat of the GRU experiment giving the highest accuracy (Fig. 6) can be explained by the fact that the target and confused activities share similar motions of the body or eyes. For example, the most confused activity for the *washing dishes* class, *washing hands*, is also rhythmic in nature, and would require a very similar physical stance to undertake. The confusions observed for *typing* were activities which might conceivably overlap with the target class. For example, *typing* might involve simultaneous *reading* of typed text, and there are also similarities to watching *TV*, with respect to the saccadic eye-movements required to observe different parts of a display. Similarly, watching of *short-form videos* was confused with *mobile admin*, *reading* and *typing*, evidently all activities involved in the use of a mobile device.

Our subject-independent classification experiments revealed a significant drop in classification accuracy when the test subjects were not present in the data used to train the network. This result suggests that the activity data captured is highly subject-dependent, perhaps because the activities were performed inconsistently, or perhaps due to strong physiological differences between subjects. This finding agrees with previously published results [28]. To overcome this and to achieve higher levels of accuracy for this task, a larger and more diverse dataset of activity data may need to be collected. Alternatively, it may be possible to adapt a network to new subjects by fine-tuning network weights using a small quantity of labelled data. Interestingly, the data for subject 4 omitted a number of activities collected by the other subjects (Table 3), and this was reflected in this subject producing the lowest mean classification accuracy in these tests. This result confirms the data shown in Fig. 6, that some activities are harder than others to distinguish, suggesting that some of the easier activities were omitted by that individual.

The results of training neural networks using eye-movement and accelerometer data independently showed that accelerometer data by itself gave improved classification accuracy over eye-movement data alone. However, the combination of both of these data channels improved the mean accuracy by 8.81% in absolute terms. When analysing the classification differences between the network trained using accelerometer data alone and that using a combination of both sources (Fig. 8), we were able to determine the activities for which eye-movements added additional discriminative power. These activities, such as *reading*, *typing* and watching *short-form videos* and *TV* predominantly require the use of the eyes, and therefore it is not surprising that the inclusion of eye-movements aided their discrimination. This result confirms that uniquely discriminative information is provided by the addition of eye-movement data. Capturing both channels of information, as made possible by the CAVA device, could enable clinicians to obtain a more detailed timeline of an individual's activities, informing their judgement on the capability for independent living, especially for activities that do not involve significant movement of the body.

8. Conclusion

The results presented in this article have demonstrated that eye-movement and accelerometer data, as captured by the CAVA device, can be used to discriminate a broad range of human movement activities. Conventional time-series classification approaches were shown to work well for this task, and the inclusion of eye-movement data significantly improved the discrimination of activities over the use of accelerometer data alone. This result makes intuitive sense, as activities such as *reading* and *typing* not only produce very distinctive excursions of the eye, but do not necessitate accompanying head movements, whereas other activities, such as *ambulating*, are more strongly associated with body movements, and eye-movements add little to aid discrimination. This important result highlights the potential benefits of using the CAVA system for activity detection compared to more conventional approaches, which typically rely on movement data alone.

A limitation of this study is the poor subject-independent test results, which confirms the results of other studies. There are two possible ways to overcome this limitation. Firstly, by the inclusion of a much larger dataset of diverse individuals undertaking longer periods of activities. Although beyond the scope of this pilot study, this is one intended avenue of our future work. A larger dataset may provide a better coverage of the feature space, allowing our networks to generalise more effectively to unseen individuals. Secondly, adaptation techniques (either semi-supervised or unsupervised) could be used, in which a short period of data from a new subject is used to fine-tune a more general model. Once we have collected a much larger dataset of data from many individuals, we will also be able to explore deeper and more complex neural network architectures and more recent neural network advancements, such as attention mechanisms.

Ultimately, we hope to expand this work to enable the automatic generation of records of daily living activities. The identification of specific activities could be used to build a timeline of events for a given day. This would enable clinicians to review a patient's function with increased context, allowing more complex, higher-level observations such as: How often is the patient eating? How frequently does the patient brush their teeth, at what times, and for how long? Such an automatically generated record could be used to replace conventional but subjective and inaccurate paper-based approaches. Beyond, this work could also aid in the generation of objective measures for disease staging, grading, and management, and for the monitoring of treatment outcomes.

CRediT authorship contribution statement

Jacob L. Newman: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zak Brook:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Stephen J. Cox:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **John S. Phillips:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization.

Ethics statement

All procedures were performed in compliance with relevant laws and institutional guidelines. The data used in this article is development data obtained from the CAVA project members themselves, and therefore informed consent is implicit.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: John Phillips, Stephen Cox, Jacob Newman has patent #GB1808649.6 issued to UEA Enterprises Limited. John Phillips, Stephen Cox, Jacob Newman has patent #17/058754 pending to UEA Enterprises Limited. John Phillips, Stephen Cox, Jacob Newman has patent #19 737 187.5 - 1122 pending to UEA Enterprises Limited. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the UK Medical Research Council under grant MR/P026265/1.

Data availability

The data relating to the article is available from the authors upon reasonable request.

References

[1] B. Walsh, C. Fogg, S. Harris, P. Roderick, S. de Lusignan, T. England, A. Clegg, S. Brailsford, S.D. Fraser, Frailty transitions and prevalence in an ageing population: longitudinal analysis of primary care data from an open cohort of adults aged 50 and over in England, 2006–2017, Age Ageing 52 (5) (2023) afad058.

[2] Y.V. Chudasama, K. Khunti, C.L. Gillies, N.N. Dhalwani, M.J. Davies, T. Yates, F. Zaccardi, Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: a longitudinal cohort study, PLoS Med. 17 (9) (2020) e1003332.

[3] L. Jönsson, A. Tate, O. Frisell, A. Wimo, The costs of dementia in Europe: an updated review and meta-analysis, Pharmacoeconomics 41 (1) (2023) 59–75.

[4] Y. Chen, P. Bandosz, G. Stoye, Y. Liu, Y. Wu, S. Lobanov-Rostovsky, E. French, M. Kivimaki, G. Livingston, J. Liao, et al., Dementia incidence trend in England and Wales, 2002–19, and projection for dementia burden to 2040: analysis of data from the English Longitudinal Study of Ageing, Lancet Public Health 8 (11) (2023) e859–e867.

[5] K.S. Frederiksen, K.L. Lanctôt, W. Weidner, J.H. Hahn-Pedersen, S. Mattke, A literature review on the burden of Alzheimer's disease on care partners, J. Alzheimer's Dis. (Preprint) (2023) 1–20.

[6] S. Rani, S.B. Dhar, A. Khajuria, D. Gupta, P.K. Jaiswal, N. Singla, M. Kaur, G. Singh, R.P. Barnwal, Advanced overview of biomarkers and techniques for early diagnosis of Alzheimer's disease, Cell Mol. Neurobiol. 43 (6) (2023) 2491–2523.

[7] G. Cipriani, S. Danti, L. Picchi, A. Nuti, M.D. Fiorino, Daily functioning and dementia, Dementia Neuropsychol. 14 (2) (2020) 93–102.

[8] C.R. Green, R.C. Mohs, J. Schmeidler, M. Aryan, K.L. Davis, Functional decline in Alzheimer's disease: a longitudinal study, J. Am. Geriatr. Soc. 41 (6) (1993) 654–661.

[9] P.F. Edemekong, D. Bomgaars, S. Sukumaran, S.B. Levy, Activities of Daily Living, StatPearls Publishing LLC, 2019.

[10] J. Cummings, J.H. Hahn-Pedersen, C.S. Eichinger, C. Freeman, A. Clark, L.R.S. Tarazona, K. Lanctôt, Exploring the relationship between patient-relevant outcomes and Alzheimer's disease progression assessed using the clinical dementia rating scale: a systematic literature review, Front. Neurol. 14 (2023) 1208802.

[11] K. Juva, M. Mäkelä, T. Erkinjuntti, R. Sulkava, R. Yukoski, J. Valvanne, R. Tilvis, Functional assessment scales in detecting dementia, Age Ageing 26 (5) (1997) 393–400.

[12] S. Katz, The index of ADL: a standardized measure of biological and psychosocial function, J. Am. Med. Assoc. 185 (1963) 914–919.

[13] C. Graf, The lawton instrumental activities of daily living scale, AJN Am. J. Nurs. 108 (4) (2008) 52–62.

[14] L.M. Shulman, I. Pretzer-Aboff, K.E. Anderson, R. Stevenson, C.G. Vaughan, A.L. Gruber-Baldini, S.G. Reich, W.J. Weiner, Subjective report versus objective measurement of activities of daily living in Parkinson's disease, Mov. Disorders 21 (6) (2006) 794–799.

[15] M. Law, L. Letts, A critical review of scales of activities of daily living, Am. J. Occup. Ther. 43 (8) (1989) 522–528.

[16] K. Buza, Activity recognition based on accelerometer data with enhanced rocket algorithm, in: 2024 IEEE 18th International Symposium on Applied Computational Intelligence and Informatics, SACI, 2024, pp. 000321–000326, http://dx.doi.org/10.1109/SACI60582.2024.10619836.

[17] S. Luqian, Z. Yuyuan, Human activity recognition using time series pattern recognition model-based on tsfresh features, in: 2021 International Wireless Communications and Mobile Computing, IWCMC, IEEE, 2021, pp. 1035–1040.

[18] A. David, R. Ramadoss, A. Ramachandran, S. Sivapatham, Activity recognition of stroke-affected people using wearable sensor, ETRI J. 45 (6) (2023) 1079–1089.

[19] J.S. Phillips, J.L. Newman, S.J. Cox, An investigation into the diagnostic accuracy, reliability, acceptability and safety of a novel device for Continuous Ambulatory Vestibular Assessment (CAVA), Sci. Rep. 9 (1) (2019) 10452.

[20] J.L. Newman, J.S. Phillips, S.J. Cox, J. FitzGerald, A. Bath, Automatic nystagmus detection and quantification in long-term continuous eye-movement data, Comput. Biol. Med. 114 (2019) 103448.

[21] D. Díaz, N. Yee, C. Daum, E. Stroulia, L. Liu, Activity classification in independent living environment with JINS MEME eyewear, in: 2018 IEEE International Conference on Pervasive Computing and Communications, PerCom, IEEE, 2018, pp. 1–9.

[22] Z. Hussain, Q.Z. Sheng, W.E. Zhang, A review and categorization of techniques on device-free human activity recognition, J. Netw. Comput. Appl. 167 (2020) 102738.

[23] A. Gupta, K. Gupta, K. Gupta, K. Gupta, 2020 International Conference on Communication and Signal Processing, ICCSP, IEEE, 2020, pp. 0915–0919.

[24] E. Bulbul, A. Cetin, I.A. Dogru, Human activity recognition using smartphones, in: 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Ismsit, IEEE, 2018, pp. 1–6.

[25] D.N. Tran, D.D. Phan, Human activities recognition in android smartphone using support vector machine, in: 2016 7th International Conference on Intelligent Systems, Modelling and Simulation, Isms, IEEE, 2016, pp. 64–68.

[26] I. Khokhlov, L. Reznik, J. Cappos, R. Bhaskar, Design of activity recognition systems with wearable sensors, in: 2018 Ieee Sensors Applications Symposium, Sas, IEEE, 2018, pp. 1–6.

[27] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, I. Korhonen, Activity classification using realistic data from wearable sensors, IEEE Trans. Inform. Technol. Biomed. 10 (1) (2006) 119–128.

[28] O.D. Lara, M.A. Labrador, A survey on human activity recognition using wearable sensors, IEEE Commun. Surv. Tutor. 15 (3) (2013) 1192–1209, http://dx.doi.org/10.1109/SURV.2012.110112.00192.

[29] N.B. Silverberg, L.M. Ryan, M.C. Carrillo, R. Sperling, R.C. Petersen, H.B. Posner, P.J. Snyder, R. Hilsabeck, M. Gallagher, J. Raber, et al., Assessment of cognition in early dementia, Alzheimer's Dementia 7 (3) (2011) e60–e76.

[30] M. Anusuya, S. Katti, Front end analysis of speech recognition: a review, Int. J. Speech Technol. 14 (2011) 99–145.

[31] A.M. John, K. Khanna, R.R. Prasad, L.G. Pillai, A review on application of fourier transform in image restoration, in: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC, IEEE, 2020, pp. 389–397.

[32] S.J. Preece, J.Y. Goulermas, L.P. Kenney, D. Howard, A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data, IEEE Trans. Biomed. Eng. 56 (3) (2008) 871–879.

[33] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[34] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[35] B. Pang, E. Nijkamp, Y.N. Wu, Deep learning with tensorflow: A review, J. Educ. Behav. Stat. 45 (2) (2020) 227–248.

[36] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (4) (2019) 917–963.

[37] M.Z.A. Pon, K.P. KK, Hyperparameter tuning of deep learning models in keras, Sparklinglight Trans. Artif. Intell. Quantum Comput. (STAIQC) 1 (1) (2021) 36–40.

[38] I. Gandin, A. Scagnetto, S. Romani, G. Barbati, Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to intensive care unit, J. Biomed. Inform. 121 (2021) 103876.

[39] S. Liu, X. Wang, Y. Xiang, H. Xu, H. Wang, B. Tang, Multi-channel fusion LSTM for medical event prediction using EHRs, J. Biomed. Inform. 127 (2022) 104011.

[40] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[41] B.C. Mateus, M. Mendes, J.T. Farinha, R. Assis, A.M. Cardoso, Comparing LSTM and GRU models to predict the condition of a pulp paper press, Energies 14 (21) (2021) 6958.

[42] H.S. Abdullah, N.H. Ali, N.A. Abdullah, Evaluating the performance and behavior of CNN, LSTM, and GRU for classification and prediction tasks, Iraqi J. Sci. 65 (3) (2024).