EEGEyeNet: a Simultaneous Electroencephalography and Eye-tracking Dataset and Benchmark for Eye Movement Prediction

Ard Kastrati[1*] **Martyna Beata Płomecka**[2*] **Damián Pascual**[1] **Lukas Wolf**[1]
Victor Gillioz[1] **Roger Wattenhofer**[1] **Nicolas Langer**[2]
[1]ETH Zurich, Switzerland
[2]University of Zurich, Switzerland
akastrati@ethz.ch, martyna.plomecka@uzh.ch

Abstract

We present a new dataset and benchmark with the goal of advancing research in the intersection of brain activities and eye movements. Our dataset, EEGEyeNet, consists of simultaneous Electroencephalography (EEG) and Eye-tracking (ET) recordings from 356 different subjects collected from three different experimental paradigms. Using this dataset, we also propose a benchmark to evaluate gaze prediction from EEG measurements. The benchmark consists of three tasks with an increasing level of difficulty: left-right, angle-amplitude and absolute position. We run extensive experiments on this benchmark in order to provide solid baselines, both based on classical machine learning models and on large neural networks. We release our complete code and data and provide a simple and easy-to-use interface to evaluate new methods.

1 Introduction

Tracking eye position is a subject of active research due to its multiple applications across different fields such as behavioural science [1], assistive technology [2], or user experience [3]. Eye tracking in combination with conventional research methods, like behavioral measures, can help assess and potentially diagnose neurological diseases such as Autism Spectrum Disorder [4], Obsessive Compulsive Disorder [5], Schizophrenia [6], Parkinson's [7], and Alzheimer's disease [8]. Additionally, eye tracking technology can be used to detect states of drowsiness [9], to support communication for locked-in patients [10], and to measure attention in marketing [11]. In the last decade, technological advances have allowed complementing eye-tracking technology with EEG — a non-invasive, minimally restrictive, and, relatively low-cost measure of mesoscale brain dynamics with high temporal resolution. Combining behavioral information gained from eye tracking with the neurophysiological markers provided by EEG enables researchers to study perceptual, attentional or cognitive processes in naturalistic situations [12]. Notably, estimating gaze position using EEG would make available gaze position information in a wide variety of studies that cannot acquire eye tracking data otherwise, either because of the unavailability of ET hardware or the lack of in-house expertise. This, in turn, could accelerate scientific discovery on human behavior and neurological and psychiatric diseases, particularly during free viewing of complex stimuli (i.e. naturalistic paradigms) and in clinical settings, in which the installation of an eye-tracker is impractical (e.g. hospital bed).

*Both authors contributed equally

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Advancing cognitive research at the intersection of brain dynamics and eye movement requires synchronized data from EEG and eye-tracking. Such data contains gaze pattern of eye movements recorded with eye- tracking as well as the neurophysiological markers provided by EEG, allowing researchers to study attention and reaction time [13], or to improve brain-computer interfaces [14]. In the same line, estimating gaze position from EEG signals is typically approached with machine learning and deep learning models [15, 16], which need significant amounts of annotated data for training. However, collecting and annotating simultaneous EEG and eye position data is time-consuming and expensive since it requires equipment and expertise for both EEG acquisition and eye-tracking. Hence, the access to concurrently recorded EEG-ET data is highly restricted, which significantly slows down progress in this field. To help bridging this gap, we release EEGEyeNet, a large dataset of EEG data synchronized with precise eye-tracking recordings. Furthermore, given the multiple benefits that EEG-based eye tracking can bring to different domains, in conjunction with our dataset, we present a benchmark for evaluating gaze estimation from EEG. This benchmark comprises three tasks of increasing difficulty and is conceived as a tool to facilitate comparable and reproducible research on gaze estimation from EEG data. We run extensive experiments on the proposed benchmark to establish baseline performance. In order to foster further research in this field we make all of our code and infrastructure available in the following site: `http://www.eegeye.net`.

To conclude, our key contributions can be summarized as:

- A dataset of high-density 128-channel EEG data synchronized with video-infrared eye tracking from 356 healthy adults that amounts to a total of more than 47 hours of recording. Along with the raw dataset we provide preprocessed data and code for preprocessing and feature extraction.
- A benchmark for gaze estimation from EEG signals that consists of three evaluation tasks with increasing difficulty. This benchmark is built on a subset of the EEGEyeNet dataset.
- An extensive experimentation that establishes baseline performance on the proposed benchmark.

2 Related Work

Gaze prediction is an active research topic with applications in human behaviour analysis [17], advertisement [18], and human-computer interaction [19], to name a few. Previous research found evidence suggesting that action selection is facilitated by attention [1]. Furthermore, [20] demonstrated the possibility of performing activity recognition from eye movements. To predict gaze location, some models use saliency maps [21, 22], while others leverage machine learning techniques to estimate gaze position from indirect data: Krafka et al. [23] use webcam images, and Son et al. [24] and LaConte et al. [25] use functional magnetic resonance imaging (fMRI). Similarly, O'Connell and Chun [26] reconstructed fixations maps, which can predict eye movement patterns, directly from fMRI data. Interestingly, our recent work indicates that it might be possible to infer gaze direction directly from EEG data [16]. Although to our knowledge no previous work has employed deep learning to estimate eye position from EEG, there are studies demonstrating that combining EEG and ET can improve performance in various tasks, as compared with a single modality. In particular, this has been reported in the vigilance estimation [27], information extraction and sentiment analysis [28], or analysing users' behaviour when performing a web search [29].

Related Datasets. There exist some openly available datasets that combine EEG and ET data. However, in comparison to EEGEyeNet, they are acquired from a smaller sample of individuals [28, 30] or using a less advanced EEG-ET setup [31, 32]. A prominent example of a multimodal neurophysiological dataset, including EEG and ET data collected from a significant sample of participants (126 individuals), was proposed by Langer et al. [33]. However, it is devised as a resource for assessing information processing in the developing brain and contains data from young participants only. This way, EEGEyeNet dataset is the first large-scale and precisely annotated EEG-ET dataset containing recording from participant across the adult lifespan (18-80 years old). In Table 1 we compare the EEGEyeNet dataset with the cited datasets in terms of the number of participants, their age, and recording/session length.

Paper/Dataset	*Participants*	*Female/Male*	*Age*	*Recording Length*	*Session Length*
ZuCo 1.0	12	5/7	25-51	x	48-72h
ZuCo 2.0	18	8/10	23-52	x	30h-54h
Notaro et al	22	11/11	18-40	7h	x
MAMEM	36	9/27	25-71	x	52h
Langer et al	126	56/70	6-44	x	378h
EEGEyeNet	356	190/166	18-80	47h	415h

Table 1: Metadata comparison of the cited datasets. Since [28, 30, 32, 33] reported only the duration of the entire experimental session (including participant's preparation, practice trials that were aimed to acquaint the participant with the experimental procedures, breaks between the subsequent experiments, etc.), we decided to distinguish the recording's length from the whole session length.

3 EEGEyeNet Dataset

In this section, we provide a detailed description of the EEGEyeNet dataset. Together with the raw data, we release two sets of preprocessed data: minimally and maximally preprocessed; as well as the preprocessing code. This way, we give users the freedom to manipulate raw data while easing the experimentation barrier by additionally providing ready-to-use clean data.

3.1 Data Acquisition

Participants. Data were recorded from 356 healthy adults. The study included 190 female and 166 male participants, of ages between 18 and 80 years. All participants gave their written informed consent before participation in the experiment and received a monetary compensation (the local currency equivalent of 50 US Dollars). The data was collected according to the principles expressed in the Declaration of Helsinki [34].

Recording setup. High-density EEG data was recorded at a sampling rate of 500 Hz, with midline central recording reference, using a 128-channel EEG Geodesic Hydrocel system. The impedance of each electrode was checked prior to each recording session and kept below 40 kOhm. Simultaneously, eye position was recorded with an infrared video-based ET EyeLink 1000 Plus from SR Research at a sampling rate of 500 Hz and an instrument spatial resolution of less than 0.01° root mean square (RMS) of the distances between successive samples. The ET was calibrated with a 9-point grid before each recording. In a validation step, the ET calibration was repeated until the error between two measurements at any point was less than 0.5°, or the average error for all points was less than 1°. Participants were seated at a distance of 68 cm from a 24-inch monitor with a resolution of 800 × 600 pixels. A stable head position was ensured with a chin rest. The illustration of the recording setup can be seen in Figure 1.

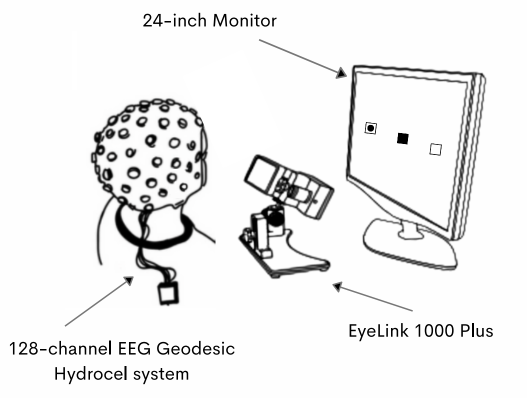

Figure 1: Recording Setup

3.2 Preprocessing

EEG data is often contaminated by artifacts produced by environmental factors, e.g., temperature, air humidity, as well as other sources of electromagnetic noise, such as line noise [35]. These artifacts interact in a complex manner with participant-related artifacts, typically reflecting unwanted physiological signals such as eye movements, eye blinks, muscular noise, heart signals or sweating, which differ from participant to participant. The resulting artifacts in the EEG data are typically more prominent than the signal of interest (i.e. brain activity). Therefore, EEG data requires preprocessing in the form of artifact cleaning or artifact correction [36]. We preprocessed the entire EEGEyeNet dataset using the openly available toolbox from Pedroni et al. [37] in two ways: minimally and maximally. Minimal preprocessing includes the detection and interpolation of bad electrodes, and filtering the data with 40 Hz high-pass filter and 0.5 Hz low-pass filter (cf. Appendix A). The difference between these two types of preprocessing is that maximal preprocessing removes a much larger number of artifacts (muscles, heart, eyes, line noise, channel noise). To do this, independent component analysis (ICA) is applied in combination with *IClabel* [38], a pre-trained classifier that estimates the probability of a component reflecting artifactual activity. If a component receives a probability estimation larger than 0.8 for any class of artifact we remove it from the data. Minimally preprocessed data includes ocular artifacts, which is expected to make the estimation of gaze position easier. On the other hand, maximal preprocessing is a state-of-the-art technique for neuroscientific applications [37] that aims to keep only neurophysiological information in the data.

After preprocessing (both minimally and maximally), the EEG and eye-tracking data were synchronized using “EYE EEG” [39] to enable EEG analyses time-locked to the onsets of relevant events depending on the experimental paradigm. Synchronization quality was ensured by comparing the trigger latencies recorded in the EEG and eye-tracker data. All synchronization errors did not exceed 2 ms.

3.3 Data Annotation

Existing literature studying eye movement generally distinguishes between three different events [40]: saccades, fixations, and blinks. **Saccades** are rapid, ballistic eye movements that instantly change the gaze position. **Fixations** are defined as time periods without saccades, and **blinks** are considered a special case of fixation, where the pupil diameter is zero. For each of the experimental paradigms described in Section 3.4, we provide annotations in the form of start and end time of each event, as well as the start and end position of saccades and the average position of fixations. See Appendix B for further details.

3.4 Experimental Paradigms

Pro- and Antisaccade. The pro- and antisaccade paradigm was designed according to the internationally standardized protocol for antisaccade testing developed by Antoniades et al. [41] and is described in detail in [42]. Each trial starts with a central fixation square. The participants are asked to focus on the center of the screen for a randomized time-period between 1 and 3.5 seconds. Subsequently, the cue (i.e. a dot) appears horizontally on the left or the right hand-side of the central fixation square. In the prosaccade trials, the participants are asked to focus their gaze on the cue as fast as possible, while in the antisaccade trials the participants are instructed to perform a saccade towards the opposite side of the cue. In both cases, the cue is shown for one second. As soon as the cue disappears, the

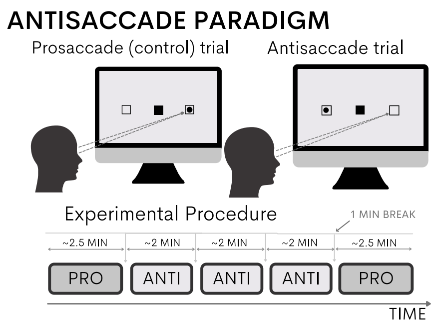

Figure 2: **Pro- and Antisaccade.** Schematic view of the experimental setup and gaze behavior during a prosaccade (left) and antisaccade (right) trial.

participants shift their focus back to the center of the screen; this is illustrated in Figure 2. Data recorded following this paradigm may be used for different research purposes, such as estimating gaze direction or examining responses to inhibition.

Large Grid. Participants are asked to fixate on a series of dots that are sequentially presented, each at one of 25 different screen positions. Unlike the others, the dot at the center of the screen appears three times, resulting in 27 trials (displayed dots) per block, each dot is displayed for 1.5 to 1.8 seconds. The positions of the dots were selected to ensure coverage of all corners of the screen as well as the center (see Figure 3). The shape of the grid and its use for eye gaze estimation follows the work from Son et al. [24]. Given that Son et al. [24] used the Large Grid paradigm for functional Magnetic Resonance Imaging (fMRI), we have adapted the length of the stimulus and the number of repetitions to our setup. To record a larger number of trials and reduce the predictability of the subsequent positions in the primary sequence of the stimulus, we use different pseudo-randomized orderings of the dots presentation, distributed in five experimental blocks, as shown in Figure 3. The entire procedure is repeated 6 times during the measurement, resulting in 810 stimuli for each participant.

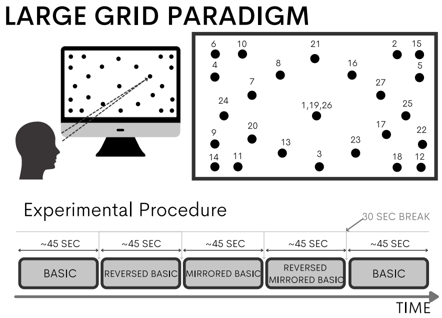

Figure 3: **Large Grid.** Schematic view of the experimental setup and stimuli location on the screen.

Visual Symbol Search. Visual Symbol Search (VSS) is a computerized version of a clinical assessment to measure processing speed (Symbol Search Subtest of the Wechsler Intelligence Scale for Children IV (WISC-IV) [43]) and the Wechsler Adult Intelligence Scale (WAIS-III) [44]). Participants are shown 15 rows at a time, where each row consists of two *target* symbols, five *search* symbols and two additional symbols that contain respectively the words "YES" and "NO". For each row, participants need to indicate by clicking with the mouse button on the "YES" or "NO" symbol, whether or not one of the two target symbols appears among the five search symbols. Each recording of the VSS paradigm takes 120 seconds with a maximum of 60 trials, where one trial corresponds to one row; in 50% of the trials one of the target symbols does appear in the search symbols and in the remaining 50% none does. Once participants finish a set of 15 rows, they press a "next page" button which displays a new set of 15 rows. Participants are instructed to solve as many rows, or trials, as possible within the given 120 seconds. Before beginning the actual recording, participants perform a training of four trials, for which they receive feedback to ensure they understand the task. No feedback is provided throughout the actual recording. Data collected according to this paradigm may be used for investigating behavioral and neurophysiological correlates of processing speed.

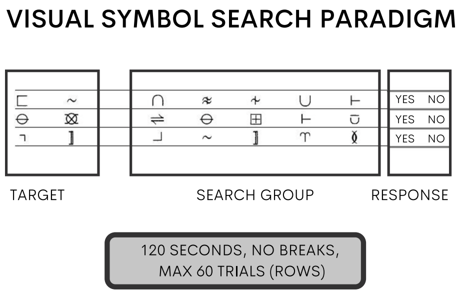

Figure 4: **Visual Symbol Search.** Three representative trials of the VSS task. For each row, the first two symbols are target symbols, the next five correspond to the search group, and the response check boxes are labeled as response.

The EEGEyeNet dataset contains data recorded following all three different experimental paradigms mentioned above. These paradigms cover typical cognitive tasks and provide a wealth of eye movement patterns.

Finally, in Table 2 we report for each paradigm the number of appearances of the three events that we extract: fixations, saccades and blinks; we report these after both minimal (min) and maximal (max) preprocessing. In Appendix C we give further details about the characteristics of each of these events. These numbers make apparent the large size of the EEGEyeNet dataset, with a total of over 47 hours of recorded events.

Paradigm	Preproc.	# Fixations	# Saccades	# Blinks	Total time
Pro- Antisac.	min	357115	358384	56179	38 h
	max	358587	359856	57991	38 h 6 mins
Large Grid	min	68075	68245	11108	7 h 52min
	max	69013	69185	11237	7 h 58 min
VSS	min	43384	43443	971	1 h 29 min
	max	43279	43339	945	1 h 28 min
Total	min	468574	470072	68258	47 h 21 min
	max	470879	472380	70173	47 h 33 min

Table 2: **Overview of EEGEyeNet Dataset.** Number of eye-tracking events (fixations, saccades, blinks) for the EEGEyeNet experimental paradigms. The difference in the number of eye events between the two preprocessing versions is due to fact that in the minimal preprocessing more events are identified as artifacts and removed from the sample.

4 Benchmark

Based on the EEGEyeNet dataset we propose a benchmark to assess EEG-based eye tracking methods. This benchmark consists of three different tasks with increasing difficulty. For each of the tasks, we provide data preparation modules that cut the data into samples of one second with a temporal resolution of $2ms$ from all 128 EEG channels; This way all samples have a shape of 500×128, i.e., 500 time points for each of the 128 EEG channels. This is also

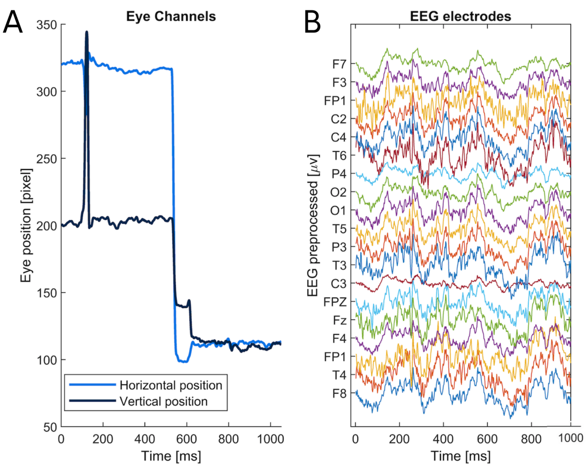

Figure 5: Each sample of EEG data has a shape of 500×128, i.e., 500 time points for each of the 128 EEG channels. On the left side (A) we can see gaze data (along XY-axes) of the one-second sample. On the right side (B) we can see a subset of the preprocessed EEG data (electrodes matching the 10–20 systems were chosen).

illustrated in Figure 5. We use exclusively minimally preprocessed EEG data as it produces better performance as compared to maximal preprocessed data. In Appendix E, we report our results for the maximally preprocessed data, as well as the results of the experiments ran on the Zuco 2.0 dataset [30]. For comparability, we also provide a stable split of the data across participants, with 70% of the participants for training, 15% for validation and 15% for test. Note that each participant's data is contained only in one of the three sets, i.e., the same participant does not appear during training and validation or test. In Table 3 we include a summary of how the data is split in each task of the benchmark.

	# Participants				# Samples			
Dataset	Total	Train	Validation	Test	Total	Train	Validation	Test
Left-Right	329	229	50	50	30842	21042	4980	4820
Angle/Amplitude	27	19	4	4	17830	12275	2836	2719
Abs. Position	27	19	4	4	21464	14706	3277	3481

Table 3: **Benchmark Data.** Detail of the data split for each task of the proposed benchmark.

4.1 Task 1: Left-right

The first and simplest task of our benchmark consists of determining the direction of the subject's gaze along the horizontal axis. This task is performed exclusively on the first experimental paradigm (*Pro- and Antisaccade*). We make sure that in all the samples provided for this task, after the cue, the participant performs a saccade towards the correct direction and then fixates on the cue; otherwise, the sample is not included in the data. Given that the antisaccade task has been viewed as an indicator of behavioral inhibition abilities [45] and that here we are interested in saccade direction, for the sake of simplicity, we use only data from the prosaccade trials of the antisaccade experimental paradigm. From the $30,842$ samples used in this task, $14,827$ correspond to the "left" label and $16,015$ to "right".

Each sample starts at the instant when a cue is shown on the screen and contains one saccade: the goal is to determine whether it is a left or a right saccade. This is a binary classification task, and therefore, performance is measured in terms of accuracy with a random baseline of 52.3%, given by the majority class. Although seemingly simple, determining the horizontal direction of the human gaze is of paramount importance in multiple real-world applications, e.g., gaze-based writing systems for people with disabilities [46].

4.2 Task 2: Angle/Amplitude

The second task of our benchmark consists of determining the angle and amplitude of a saccade and is performed on data from the *Large Grid* paradigm. The one-second samples in this task contain a saccade onset in the middle of the sample. This way, each sample contains a complete saccade as well as a part of the preceding and succeeding fixations. During the data preparation for this task, we remove samples with fixations that are shorter than $150ms$ and with saccades with an amplitude of less than 1° [47].

Given a sample, the task of the model is to regress the two target values, i.e., angle and amplitude of the relative change of the gaze position during the saccade. The evaluation metric for this task is the square root of the mean squared error (RMSE) for the angle (in radians along the shortest path of the unit circle) and amplitude (in millimeters) separately. The naive baseline is given by the mean angle and amplitude in the training set and amounts to 1.90 RMSE radians for the angle and 74.7 RMSE mm for the amplitude. This task is significantly harder than the Left-Right task and aims at serving as an intermediate step for the development of a purely EEG-based ET. Despite its difficulty, there is evidence that EEG recordings contain angle information [48].

4.3 Task 3: Absolute Position

Finally, we propose the task of determining the absolute position of the subject's gaze in the screen, described in terms of XY-coordinates. Again, this task is performed on the *Large Grid* paradigm. We provide samples of one second during which the participant is performing only one fixation. The data preparation module ensures that in this time window there is no other event happening, i.e., the participant performs only a fixation. However, we note that in order to estimate the current gaze position we expect past information, e.g., the previous gaze position, to be helpful. For this reason, we also encourage experimentation on different ways of processing and cutting the full EEG recordings provided in the dataset.

Performance is measured as the euclidean distance in millimeters between the actual and the estimated gaze position in the XY-coordinate plane. Random performance is again calculated as the mean position across the training set and corresponds to a distance of 123.3 mm. This is the hardest task in the proposed benchmark and aims at simulating a purely EEG-based eye-tracker. We expect this task to help in the development of general methods to further improve gaze estimation systems as well as in setting an upper bound on the reach of EEG-based eye-tracking.

5 Baselines

We run extensive experiments on the proposed benchmark in order to provide baselines of different complexity. In our repository we provide an intuitive interface to reproduce our results and to use the methods presented here as a starting point for further research. We consider both models based on classical machine learning as well as large neural networks.

5.1 Models

Machine Learning. These models operate on features extracted from the data rather than on the raw data. Therefore, in a feature extraction step, we apply a band-pass filter in the alpha band [$8 - 13$ Hz] on the continuous EEG signals across the entire trial. The choice of the alpha band is motivated by growing evidence suggesting that alpha activity is integral to spatial attention, and therefore plays a central role in covert orienting in a range of paradigms [49]. After band-passing the signal, the Hilbert transform was applied, resulting in a complex time series from which we extract phase and amplitude. Using the resulting features, we experiment with different models. *Left-Right* is a classification task while *Angle/Amplitude* and *Absolute Position* are regression tasks and thus, some of the considered models can be applied only to either of those two types of tasks. In particular, the classification-only models that we study are Gaussian Naive Bayes (NB) Linear Support Vector Classification (SVC) and Radial Basis Function (RBF) kernel SVC, whereas the regression-only models are Linear, Ridge and Lasso Regression as well as Elastic Net and RBF Support Vector Machine for regression (SVR). Furthermore, we use K-Nearest Neighbours (KNN) and four tree-based models, Random Forest, Gradient Boost, AdaBoost and XGBoost, which can be used for both, classification and regression. In all cases we use the implementation from the Sklearn library [50], for detailed model hyperparameters see Appendix D.

Deep Learning. We evaluate five state-of-the-art architectures for time series on the proposed benchmark: a standard one-dimensional convolutional neural network (CNN), a CNN with pyramidal shape, the EEGNet model by Lawhern et al. [51], an InceptionTime model [52], and an Xception model [53]. All of these models use convolutions as the primary operation (see Appendix D for architectural details). We tune the learning rate and other hyperparameters on the validation set of the *Left-Right* task and use those values in all the reported results (cf. Appendix D). We use binary cross entropy loss to train the models for *Left-Right*, and mean square error (MSE) loss for the other two tasks; in all cases we use the Adam optimizer [54] and early stopping on the validation sets.

5.2 Results

In Table 4 we provide the results of our evaluation for each of the models considered and for each of the three tasks. We tune the learning rate of the different deep learning models on the validation set of the *Left-Right* task and use the same rate in the other two tasks, i.e., $1e^{-4}$. We run 5 times each experiment and report mean performance and standard deviation.

Model	Left-Right Accuracy	Angle/Amplitude Angle RMSE	Angle/Amplitude Amp. RMSE	Abs. Position RMSE
KNN	90.7 ±0	1.26 ±0	59.3 ±0	119.7 ±0
GaussianNB	87.7 ±0	-	-	-
LinearSVC	92.0 ±0	-	-	-
RBF SVC/SVR	89.4 ±0	1.88 ±0	75.9 ±0	123 ±0
Linear Regression	-	1.39 ±0	64.6 ±0	118.3 ±0
Ridge Regression	-	1.39 ±0	64.2 ±0	118.2 ±0
Lasso Regression	-	1.38 ±0	63.9 ±0	118 ±0
Elastic Net	-	1.38 ±0	63.6 ±0	118.1 ±0
Random Forest	96.5 ±0	1.09 ±0.01	59.8 ±0.1	116.7 ±0.1
Gradient Boost	97.3 ±0.1	1.11 ±0.01	60 ±0.1	117 ±0.1
AdaBoost	96.3 ±0	1.43 ±0.01	65 ±0.1	119.4 ±0.1
XGBoost	97.9 ±0	1.11 ±0	61.3 ±0	118 ±0
CNN	98.3 ±0.5	**0.33** ±0.05	32 ±3.6	**70.2** ±1.1
PyramidalCNN	98.5 ±0.2	0.34 ±0.04	**30.7** ±1	73.6 ±1.9
EEGNet	98.6 ±0.1	0.70 ±0.08	46 ±5.2	81.7 ±1.0
InceptionTime	97.9 ±1.1	0.44 ±0.16	43.6 ±21.85	70.8 ±0.8
Xception	**98.8** ±0.1	0.47 ±0.28	32.2 ±1.9	78.7 ±1.6
Naive Baseline	52.3	1.90	74.7	123.3

Table 4: **Results.** Mean and standard deviation of 5 runs of the considered models on the three benchmark tasks. *Angle* is measured in radians, *Amplitude* and *Abs. Position* in mm.

Left-Right. We see that classical machine learning models achieve high performance in this task, much above the naive baseline of 52.26%. In particular, tree-based models reach a performance of over 96%, which confirms Left-Right as the easiest task in the proposed benchmark. Notably, although classical models obtain high scores, all the deep learning models reach a performance consistently higher, with an accuracy of over 98% in all cases (except for InceptionTime, 97.9%). This shows that despite the high performance, differences in performance can still be noticeable.

Angle/Amplitude. The results in Table 4 clearly show that this task is harder than *Left-Right*. Except for RBF SVC/SVR, which performs close to random, all classical machine learning models reach a similar performance for the estimation of both amplitude and angle. This result is above the naive baseline although not by a big margin. Tree-based models perform the best among classical statistical models, however, they are clearly inferior to deep learning models in this task. Somewhat surprisingly, among the deep learning models the simplest ones, i.e., CNN and PyramidalCNN perform the best, with an RMSE of 0.33 and 0.34 radians in angle estimation, and 32 and 30.7 mm in amplitude estimation, respectively. Overall, we see that there is still a considerable gap between the best results reported here and ideal performance. Our results constitute a baseline that should orient future work on estimating angle and amplitude of saccades from EEG data.

Absolute Position. Finally, we see in the last column of Table 4 the results of our experimentation on absolute position estimation. We see that classical models generally fail in this task, with performances very close to the naive baseline. On the other hand, deep learning models reach an euclidean distance with respect to the true location in the range of 70 to 80 mm. Although these results are far from ideal performance, the considerable gap with the naive baseline shows that EEG-based eye tracking can potentially be achieved to an acceptable degree of accuracy. Our results, set the baseline for this task in 70.2 mm, as reached by CNN, again one of the simpler deep learning models. However, as explained in Section 4.3, exploiting previous information is likely to improve performance, and thus, it would be interesting to see how sequence models, such as Recurrent Neural Networks or

Transformers would perform in this task. There is a lot of room for improvement and we hope that this task will help future work in advancing EEG-based eye tracking.

In summary, our evaluation reveals that deep learning models are superior to other statistical techniques in estimating gaze position from EEG data. Although this is not surprising, given the complexity of the task and the larger expressive capacity of neural networks, it confirms that EEGEyeNet is a valuable resource for developing large neural models. We expect that future work will surpass our scores advancing EEG-based eye tracking.

6 Discussion and Future Work

The dataset and benchmark presented in this work provide an approachable framework to conduct research in the intersection of brain activity and eye movement. We will actively maintain and continuously extend both the dataset and the benchmark with further measurements from more participants and new experimental paradigms. We acknowledge as current limitations the relatively small number of participants recorded for the *Large Grid* paradigm and the fact that the test sets of our benchmark are publicly available. We will address both points by (1) recording more data and (2) building an automatic evaluation service that keeps the test set hidden from the users and includes a leaderboard.

To facilitate extensive use of EEGEyeNet for various research purposes, we provide in our repository data preparation tools with an easy-to-use interface where users can define their own benchmarking tasks or extract other information from the dataset. In particular, the user can specify whether some recording blocks from the experimental paradigm should be ignored, which events to extract from the data and how the data should be preprocessed. Additionally, the user can also decide whether feature extraction should be performed or not. As with the other resources presented in this work, the data preparation module is in continuous development. We plan to adapt this software tool in the future according to the users' needs. Overall, we expect that EEGEyeNet will become a central resource for a broad range of EEG and Eye-Tracking related research, specifically:

Research in Cognitive Area. EEGEyeNet's rich structure and high-density coverage of EEG and Eye-Tracking data may help advance other areas that study the combination of gaze position and brain activity to identify variations in attention, arousal and participant's compliance with the task demands. Moreover, the behavioral information gained from eye tracking with the high temporal resolution and neurophysiological markers provided by EEG enables research of the perceptual, attentional, cognitive processes.

Benchmarking. We expect the high quality, diversity and large scale of the EEGEyeNet dataset to be leveraged, in order to include new tasks in the proposed benchmark, as well as to build benchmarks for related domains. In particular, we plan to include segmentation tasks that evaluate the ability of a given model to detect and distinguish events such as fixations or saccades.

7 Conclusion

Recording eye-tracking data is a complex and expensive process that requires specialized hardware, trained operators and participants' consent. Collecting such data in combination with EEG adds an additional level of complexity to the data acquisition process. Therefore, behavioral research studying the combination of brain activity and eye movements is typically restricted by the lack of appropriate data. In this work, we have introduced EEGEyeNet, a large dataset of EEG and eye tracking data, with the view of making basic cognitive neuroscience research more approachable. Furthermore, given the potential benefits that EEG-based eye-tracking can bring in different domains, we have proposed a benchmark to facilitate the development of new methods tackling this challenge. Our experiments on this benchmark show that deep learning models perform better than classical statistical models. This confirms that the amount of data contained in EEGEyeNet is large enough to reliably train large neural models, which is a promising direction for further developing EEG-based eye tracking.

References

[1] Maria K Eckstein, Belén Guerra-Carrillo, Alison T Miller Singley, and Silvia A Bunge. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental cognitive neuroscience*, 25:69–91, 2017.

[2] Radiah Rivu, Yasmeen Abdrabou, Ken Pfeuffer, Augusto Esteves, Stefanie Meitner, and Florian Alt. Stare: Gaze-assisted face-to-face communication in augmented reality. ETRA '20 Adjunct. Association for Computing Machinery, 2020.

[3] Jennifer Romano Bergstrom and Andrew Schall. *Eye tracking in user experience design*. Elsevier, 2014.

[4] Terje Falck-Ytter, Sven Bölte, and Gustaf Gredebäck. Eye tracking in early autism research. *Journal of neurodevelopmental disorders*, 5(1):1–13, 2013.

[5] Maria C Bradley, Donncha Hanna, Paul Wilson, Gareth Scott, Paul Quinn, and Kevin FW Dyer. Obsessive–compulsive symptoms and attentional bias: An eye-tracking methodology. *Journal of behavior therapy and experimental psychiatry*, 50:303–308, 2016.

[6] Philip S Holzman, Leonard R Proctor, and Dominic W Hughes. Eye-tracking patterns in schizophrenia. *Science*, 181(4095):179–181, 1973.

[7] Hiroshi Shibasaki, Sadatoshi Tsuji, and Yoshigoro Kuroiwa. Oculomotor abnormalities in parkinson's disease. *Archives of Neurology*, 36(6):360–364, 1979.

[8] J Thomas Hutton, JA Nagel, and Ruth B Loewenson. Eye tracking dysfunction in alzheimer-type dementia. *Neurology*, 34(1):99–99, 1984.

[9] Mandalapu Sarada Devi and Preeti R Bajaj. Driver fatigue detection based on eye tracking. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 649–652. IEEE, 2008.

[10] Ziad O Abu-Faraj, Maya J Mashaalany, Habib C Bou Sleiman, Jean-Louis D Heneine, and Waleed M Al Katergi. Design and development of a low-cost eye tracking system for the rehabilitation of the completely locked-in patient. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4905–4908. IEEE, 2006.

[11] Michel Wedel and Rik Pieters. *Eye tracking for visual marketing*. Now Publishers Inc, 2008.

[12] Peter König, Michael Plöchl, and Jose Pablo Ossandón. Combining eeg and eye tracking: Identification, characterization and correction of eye movement artifacts in electroencephalographic data. *Biomedical Engineering/Biomedizinische Technik*, 57 (SI-1-Track-F):392–392, 2012.

[13] Michael Plöchl, José Pablo Ossandón, and Peter König. Combining eeg and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in human neuroscience*, 6:278, 2012.

[14] Xinyao Ma, Zhaolin Yao, Yijun Wang, Weihua Pei, and Hongda Chen. Combining brain-computer interface and eye tracking for high-speed text entry in virtual reality. In *23rd International Conference on Intelligent User Interfaces*, pages 263–267, 2018.

[15] Sergei L Shishkin, Yuri O Nuzhdin, Evgeny P Svirin, Alexander G Trofimov, Anastasia A Fedorova, Bogdan L Kozyrskiy, and Boris M Velichkovsky. Eeg negativity in fixations used for gaze-based control: Toward converting intentions into actions with an eye-brain-computer interface. *Frontiers in neuroscience*, 10:528, 2016.

[16] Ard Kastrati, Martyna Beata Plomecka, Roger Wattenhofer, and Nicolas Langer. Using deep learning to classify saccade direction from brain activity. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '21 Short Papers. Association for Computing Machinery, 2021.

[17] Alexandros André Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[18] Genki Okada, Kenta Masui, and Norimichi Tsumura. Advertisement effectiveness estimation based on crowdsourced multimodal affective responses. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*, pages 1263–1271, 2018.

[19] Päivi Majaranta and Andreas Bulling. Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing*, pages 39–65. Springer, 2014.

[20] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Tröster. Eye movement analysis for activity recognition using electrooculography. *IEEE transactions on pattern analysis and machine intelligence*, 33(4):741–753, 2010.

[21] Ryoichi Nakashima, Yu Fang, Yasuhiro Hatori, Akinori Hiratani, Kazumichi Matsumiya, Ichiro Kuriki, and Satoshi Shioiri. Saliency-based gaze prediction based on head direction. *Vision research*, 117:59–66, 2015.

[22] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[23] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[24] Jake Son, Lei Ai, Ryan Lim, Ting Xu, Stanley Colcombe, Alexandre Rosa Franco, Jessica Cloud, Stephen LaConte, Jonathan Lisinski, Arno Klein, et al. Evaluating fmri-based estimation of eye gaze during naturalistic viewing. *Cerebral Cortex*, 30(3):1171–1184, 2020.

[25] S LaConte, S Peltier, K Heberlein, and X Hu. Predictive eye estimation regression (peer) for simultaneous eye tracking and fmri. In *Proc. Intl. Soc. Magn. Reson. Med*, volume 2808, page 13, 2006.

[26] Thomas P O'Connell and Marvin M Chun. Predicting eye movement patterns from fmri responses to natural scenes. *Nature communications*, 9(1):1–15, 2018.

[27] Wei-Long Zheng and Bao-Liang Lu. A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of neural engineering*, 14(2):026017, 2017.

[28] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

[29] Gino Slanzi, Jorge A Balazs, and Juan D Velásquez. Combining eye tracking, pupil dilation and eeg analysis for predicting web users click intention. *Information Fusion*, 35:51–57, 2017.

[30] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *CoRR*, abs/1912.00903, 2019. URL `http://arxiv.org/abs/1912.00903`.

[31] Gina M Notaro and Solomon G Diamond. Simultaneous eeg, eye-tracking, behavioral, and screen-capture data during online german language learning. *Data in brief*, 21: 1937–1943, 2018.

[32] Spiros Nikolopoulos, Panagiotis C Petrantonakis, Kostas Georgiadis, Fotis Kalaganis, Georgios Liaros, Ioulietta Lazarou, Katerina Adam, Anastasios Papazoglou-Chalikias, Elisavet Chatzilari, Vangelis P Oikonomou, et al. A multimodal dataset for authoring and editing multimedia content: The mamem project. *Data in brief*, 15:1048–1056, 2017.

[33] Nicolas Langer, Erica J Ho, Lindsay M Alexander, Helen Y Xu, Renee K Jozanovic, Simon Henin, Agustin Petroni, Samantha Cohen, Enitan T Marcelle, Lucas C Parra, et al. A resource for assessing information processing in the developing brain using eeg and eye tracking. *Scientific data*, 4(1):1–20, 2017.

[34] World Medical Association et al. World medical association declaration of helsinki: ethical principles for medical research involving human subjects. *Jama*, 310(20):2191–2194, 2013.

[35] Emily S Kappenman and Steven J Luck. The effects of electrode impedance on data quality and statistical significance in erp recordings. *Psychophysiology*, 47(5):888–904, 2010.

[36] Andreas Keil, Stefan Debener, Gabriele Gratton, Markus Junghöfer, Emily S Kappenman, Steven J Luck, Phan Luu, Gregory A Miller, and Cindy M Yee. Committee report: publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, 51(1):1–21, 2014.

[37] Andreas Pedroni, Amirreza Bahreini, and Nicolas Langer. Automagic: Standardized preprocessing of big eeg data. *Neuroimage*, 200:460–473, 2019.

[38] Luca Pion-Tonachini, Ken Kreutz-Delgado, and Scott Makeig. Iclabel: An automated electroencephalographic independent component classifier, dataset, and website. *NeuroImage*, 198:181–197, 2019.

[39] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. Coregistration of eye movements and eeg in natural reading: analyses and review. *Journal of experimental psychology: General*, 140(4):552, 2011.

[40] Miika Toivanen, Kati Pettersson, and Kristian Lukander. A probabilistic real-time algorithm for detecting blinks, saccades, and fixations from eog data. *Journal of Eye Movement Research*, 8(2), 2015.

[41] Chrystalina Antoniades, Ulrich Ettinger, Bertrand Gaymard, Iain Gilchrist, Arni Kristjánsson, Christopher Kennard, R John Leigh, Imran Noorani, Pierre Pouget, Nikolaos Smyrnis, et al. An internationally standardised antisaccade protocol. *Vision research*, 84:1–5, 2013.

[42] Martyna Beata Płomecka, Zofia Barańczuk-Turska, Christian Pfeiffer, and Nicolas Langer. Aging effects and test–retest reliability of inhibitory control for saccadic eye movements. *Eneuro*, 7(5), 2020.

[43] David Wechsler and Habuku Kodama. *Wechsler intelligence scale for children*, volume 1. Psychological corporation New York, 1949.

[44] David Wechsler. Wechsler adult intelligence scale–. *Archives of Clinical Neuropsychology*, 1955.

[45] Wolfgang Klimesch, Paul Sauseng, and Simon Hanslmayr. Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, 53(1):63–88, 2007.

[46] Richard Bates, Michael Donegan, Howell O Istance, John Paulin Hansen, and K-J Räihä. Introducing cogain: communication by gaze interaction. *Universal Access in the Information Society*, 6(2):159–166, 2007.

[47] Susana Martinez-Conde, Jorge Otero-Millan, and Stephen L Macknik. The impact of microsaccades on vision: towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, 14(2):83–96, 2013.

[48] L'uboš Hládek, Bernd Porr, and W Owen Brimijoin. Real-time estimation of horizontal gaze angle by saccade integration using in-ear electrooculography. *Plos one*, 13(1): e0190420, 2018.

[49] Joshua J Foster, David W Sutterer, John T Serences, Edward K Vogel, and Edward Awh. Alpha-band oscillations enable spatially and temporally resolved tracking of covert spatial attention. *Psychological science*, 28(7):929–941, 2017.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[51] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[52] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[53] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[54] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[55] Arnaud Delorme and Scott Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[56] Alain de Cheveigné. Zapline: A simple and effective method to remove power line artifacts. *Neuroimage*, 207:116356, 2020.

[57] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[58] Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F Schmidt, Jonathan Weber, Geoffrey I Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020.

[59] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[60] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[61] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Author statement

Hereby we confirm that we bear all responsibility in case of violation of rights, etc., and confirmation of the data license. All data are de-identified and participants gave permission for their data to be openly shared as part of the informed consent process.

We release our code used for our experiments, data collection and data preparation at the following repository : https://github.com/ardkastrati/EEGEyeNet and our dataset at https://doi.org/10.17605/OSF.IO/KTV7M

Acknowledgments

We thank Tzvetan Popov for useful discussions related to this project and Marta Marciniak for the language editing, and proofreading. Dataset collection and maintenance is generously supported by the Velux Foundation and Swiss National Science Foundation.

We thank the participants for taking the time to be in the study, as well as their willingness to have their data shared with the scientific community.