Pacific Symposium on Biocomputing 27:242-253(2022)

Predicting Visuo-Motor Diseases From Eye Tracking Data

Kailas Vodrahalli[1,*], Maciej Filipkowski[2], Tiffany Chen[2], James Zou[1,3,*,†], and Yaping Joyce Liao[2,*,†]

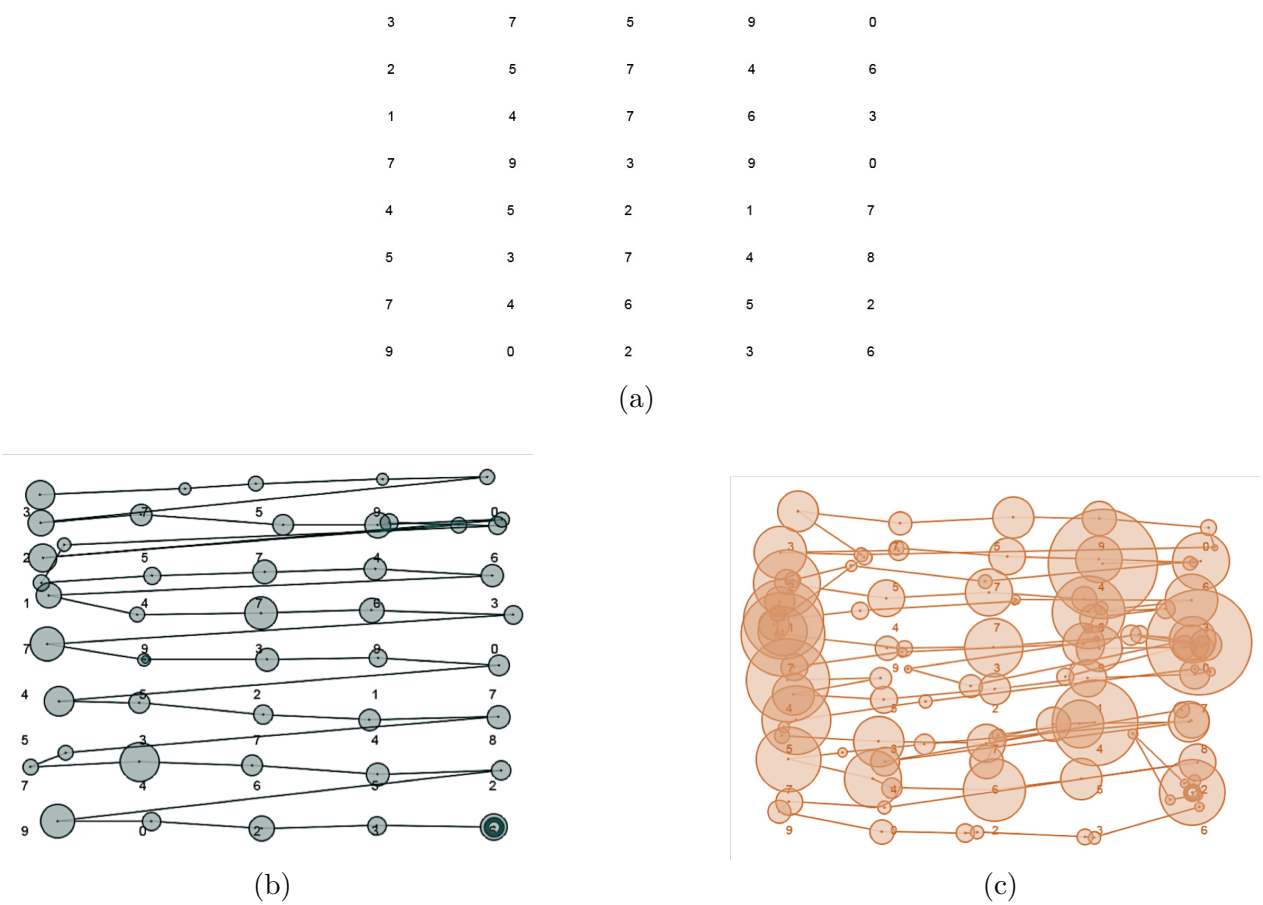
[1]*Electrical Engineering, Stanford University, Stanford, CA 94305, USA*
[2]*Department of Ophthalmology, Stanford University School of Medicine, Stanford, CA 94305, USA*
[3]*Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA*
[*]*Correspondence: kailasv@stanford.edu, jamesz@stanford.edu, yjliao@stanford.edu*
[†]*: Co-senior authors*

Eye tracking, or oculography, provides insight into where a person is looking. Recent advances in camera technology and machine learning have enabled prevalent devices like smart-phones to track gaze and visuo-motor behavior at near clinical-quality resolution. A critical gap in using oculography to diagnose visuo-motor dysfunction on a large scale is in the design of visual task paradigms, algorithms for diagnosis, and sufficiently large datasets. In this study, we used a 500 Hz infrared oculography dataset in healthy controls and patients with various neurological diseases causing visuo-motor abnormality due to eye movement disorder or vision loss. We used novel visuo-motor tasks involving rapid reading of 40 single-digit numbers per page and developed a machine learning algorithm for predicting disease state. We show that oculography data acquired while a person reads one page of 40 single-digit numbers (15-30 seconds duration) is predictive of of visuo-motor dysfunction (ROC-AUC = 0.973). Remarkably, we also find that short recordings of about 2.5 seconds (6-12× reduction in time) are sufficient for disease detection (ROC-AUC = 0.831). We identify which tasks are most informative for identifying visuo-motor dysfunction (those with the most visual crowding), and more specifically, which aspects of the task are most predictive (the recording segments where gaze moves vertically across lines). In addition to segregating disease and controls, our novel visuo-motor paradigms can discriminate among diseases impacting eye movement, diseases associated with vision loss, and healthy controls (81% accuracy compared with baseline of 33%).

Keywords: Ophthalmology; Eye-tracking; Machine Learning.

1. Introduction

Video infrared eye trackers are noninvasive ways of assessing visuo-motor abnormality in patients with diseases that impact eye or brain function. Common diseases that affect the eye-brain network include stroke, multiple sclerosis, brain tumors, traumatic brain injury, and Parkinson's disease (PD).[1–5] We used infrared oculography to train a machine learning-based algorithm that can assess visuo-motor behavior while patients complete a variety of tasks



(a)

(b)

(c)

Fig. 1. Example task and eye tracking recordings. (a) Example task: the 8 rows and 5 columns of single digit numbers are to be read left-to-right, top-to-bottom. (b,c) Recordings overlaid with reading task. Circles represent fixations; lines connect temporally adjacent fixations. Circle radii correspond to fixation duration. (b) Healthy control participant. (c) Patient with Parkinson's disease.

modified from the King-Devick (KD) test.[6] These visuo-motor tasks consist of 9 pages of 40 single-digit numbers, and 3 pages of 40 single-digit numbers written as words, which are read aloud from left-to-right or right-to-left as quickly as possible from a computer screen. The single-digit numbers are regularly or irregularly space, and some pages have greater visual crowding while others have a line that guides gaze from one number to next across the line of reading. Each page of the task takes about 15-30 seconds to complete depending on the reading speed. We show an example page and corresponding eye tracking output in Figure 1.

The raw eye tracking output contains the XY positions of each eye across time (typically sampled at rates >= 500 Hz). This output is preprocessed into a time series of *fixations* and *saccades*. A fixation is an eye movement when the gaze is stabilized on a single target. Fixations are when people actually "see." Saccades are rapid movements that move gaze from one target to the next. People are "blind" during saccades because of the rapid speed of eye movement. Simple features like the average duration of a fixation or the number of saccades in between fixations as well as more complex features like the rate of regressive saccades (saccades that

move opposite to the reading direction) can be predictive of disease state.[4,5]

Our contributions In this work, we leverage machine learning to create an end-to-end disease detection pipeline based on eye tracking data. Though we are limited by the amount of available data, we show that oculography data captured while a person reads one page of 40 single-digit numbers (duration 15-30 seconds) is predictive of visuo-motor dysfunction (ROC-AUC = 0.973). Remarkably, we also find that short recordings of about 2.5 seconds are sufficient to identify disease state (ROC-AUC = 0.831). These short recordings represent a 6-12× more efficient diagnosis system. In addition to predicting disease state, we are able to differentiate between 3 disease groups (those impacting eye movement, vision loss, and healthy controls) with 81% accuracy compared to random-guessing baseline of 33%.

Furthermore, we analyze which tasks are most predictive of visuo-motor dysfunction. We find that tasks with higher degrees of visual crowding and, in particular, recording segments where a person moves across lines while reading are the most predictive. These findings are important as they will enable us to better design tasks to identify disease state.

Limitations While our study is already one of the largest ones of its type, we are limited by the amount of data. Though our results appear promising, our model needs further validation to ensure robustness. We are actively collecting a larger dataset from diverse patients for additional validation.

2. Related Works

Diagnostic approaches in ophthalmology Visuo-motor abnormalities are typically diagnosed through a clinical neuro-ophthalmic examination.[7] Depending on the exam results, additional visual function assessment and ophthalmic imaging may be required.[8] For example, brain MRIs are a common assessment for confirming a diagnosis; other disease-specific assessments also exist.[9] The oculography recordings we use are not typically part of a clinical assessment; however, oculography is a more simple test to conduct in practice and, as we demonstrate, it is amenable to machine learning approaches for accurate patient diagnosis.

Machine learning for oculography There are a few works in recent years that use eye tracking data for disease prediction. There are two general approaches:

1. *Deep learning for end-to-end disease prediction:* Bautista et al. use a deep model to embed raw gaze measurements for use in biometrics.[10] Convolutional neural networks (CNNs) have also been leveraged to process heatmap-encoded eye tracking recording images.[11,12] Recurrent neural networks (RNNs) are used in Mao et al. to generate features that are subsequently input to a random forest model. Their work is most similar to ours, though we use a CNN model and work with discrete chunks rather than embedding the entire recording with an RNN. We discuss these differences more in Section 3.
2. *Feature engineering and a non-deep learning classifier:* This type of work goes back several decades, where researchers have identified key features associated with various diseases.[1,2,4,5,13] More recent studies use these known feature associations together with ad-

ditional engineered features to automatically predict disease state. For example, Daley et al. predict sleep deprivation using hand-picked features and compare a variety of classic ML models including support vector machines and decision trees.[14] Other works take similar approaches where feature engineering together with classic ML models are used for differentiating control and diseased states.[15,16]

Our work extends the prior work in a few ways. First, we are able to differentiate between up to 5 disease states (including controls), while most prior work only focuses on a binary task: differentiating between healthy and a disease of interest. Second, we show it is possible to accurately differentiate control from disease based on only a few seconds of data. Previous work typically uses longer tasks comparable to our task-level predictions (i.e., > 15 seconds). Lastly, we are able to provide insights into what tasks are the most useful for disease detection, which will better enable future algorithm development.

3. Methods

3.1. *Infrared oculography and data collection*

Table 1. Dataset statistics by disease group. "Number of recordings" is the total number of individual tasks across all patients. Each patient completed up to 12 unique tasks.

Disease group	Number of Patients	Number of recordings	Average Age	Percent Female
Control	35	331	46.2	57.1%
Vision Loss	16	162	59.3	62.5%
Parkinson's Disease	42	399	69.4	50.0%
Cerebellar Ataxia	13	147	50.7	53.8%
Downbeat Nystagmus	21	200	61.8	38.1%
Total	127	1239	58.8	52.0%

We performed a prospective, case control study at Stanford University from 2015-2020 using a 500 Hz 2-dimensional binocular infrared video oculography instrument (RED500, SensoMotoric Instruments, Germany). The study adhered to the tenets of the Declaration of Helsinki for human subject research and was approved by the Stanford Institutional Review Board (IRB). The study was also compliant with Health Insurance Probability and Accountability Act (HIPAA).

Participants sat approximately 70 cm in front of a computer monitor (dimensions: 55.9 cm by 33.0 cm, 22 inches by 13 inches, 1680 px by 1050 px), where the task paradigms were displayed. Participants were instructed to sit upright with both eyes open for binocular recording. Participants were instructed to minimize all head movements and stabilize their posture to avoid noise and low accuracy of eye tracking. They were repositioned and stabilized by a test administrator as needed. A chin rest was not used to ease testing because number reading was completed out loud. After set up was complete, participants completed a 9-point calibration with cardinal points of regard to accurately measure eye movements and

gaze. Validation steps using fixation points were included to confirm accurate, high quality calibration and minimize tracking errors. Horizontal and vertical deviations were measured using gaze points with the known calibration target locations. An administrator was present to troubleshoot technical issues to ensure the best quality of recording and to minimize errors for optimal tracking ratios. Infrared oculography ensured participants read all numbers and words correctly and did not skip lines. Participants were permitted to wear their reading glasses or contacts during the study as long as they did not interfere with eye movement recording. Recordings were stopped or excluded if quality of recording was poor.

Ground-truth labeling of disease status and basic demographic information including age and gender were recorded for each individual. A summary of the dataset statistics after prepossessing is shown in Table 1. We consider 4 disease states:

1. **Vision loss**: Acquired diseases of the brain such as stroke, trauma, or multiple sclerosis, that interrupts the retino-geniculo-cortical visual pathway resulting in left or right homonymous hemianopsia. Patients experience difficulty seeing and interpreting the left or right side of the visual field of vision in each eye.[17]
2. **Cerebellar ataxia** (Ataxia): Acquired neurological disorders due to impairment of the cerebellum and their pathways. This condition affects the precise control and coordination of eye movement and extraocular muscles, leading to oscillopsia (jumping vision), dizziness, and visual disability.[18]
3. **Parkinson's disease** (PD): The most common progressive neurodegenerative disorder affecting motor function. It is characterized by loss of dopaminergic and other neurons, leading to impairment of visuo-motor function and higher order cognitive abilities.[3]
4. **Downbeat nystagmus** (DBN): A disorder characterized by oscillopsia (jumping vision) due to impaired connection between the brainstem and cerebellum. Patients with DBN have difficulty holding the eyes still due to a characteristic continuous upward drift in gaze that is corrected by downward saccades. The continuous oscillations of the eyes occur at 0.5 to 3 Hz.[19]

In our analysis, we will also aggregate PD, Ataxia, and DBN into a single group, "Eye movement disorder." These three disease states share common symptoms in difficulty with motor function of the eye, and so we expect more similarities in the eye tracking recordings between these states than between healthy controls or patients with vision loss.

3.2. *Algorithm*

The data was preprocessed using commercial software BeGaze (SensoMotoric Instruments) to identify fixations and saccades. We show an overview of our algorithm in Figure 2. When training our models, we randomly split our data with a 75%-25% split for train and test data. Data is split at the patient level to avoid data leakage. Our algorithm processes short temporal "chunks" of recording data and then aggregates chunks across a recording. A chunk is defined as a short, contiguous widow of time in a recording and includes all fixations and saccades that occur inside that window.

There are three reasons we start at the chunk level. (1) We can train more complex models

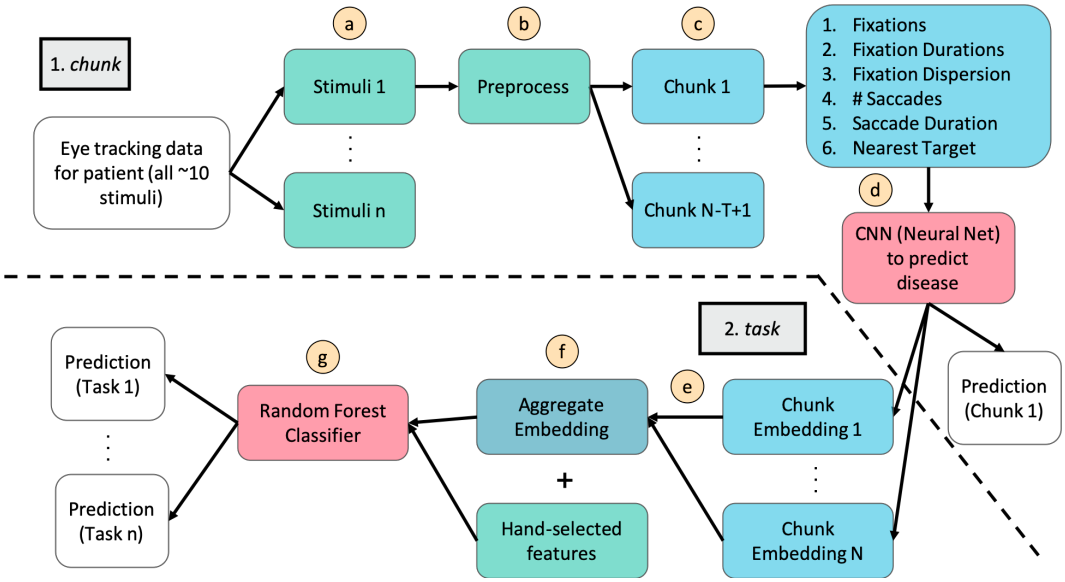


Fig. 2. Algorithm overview. (a) Retrieve patient recordings. (b) Normalize and align recordings. (c) Split recordings into roughly 3-second-long chunks. (d) Generate chunk-embeddings using a CNN trained to classify disease. (e) Aggregate embeddings produced by the chunk-classifier. (f) Concatenate aggregated embedding with task-level features. (g) Generate task-level prediction.

more reliably at the chunk level as we have more datapoints here (Table 2). (2) Chunk-level performance gives insight into what information is useful within a task and enables us to design more informative tasks for disease detection. (3) We believe short recordings are sufficient for disease prediction; an algorithm relying on less data is beneficial for widespread usage.

3.2.1. *Chunk-level classifier*

Table 2. Number of chunks per disease group. On average, there are around 60 chunks per task.

	Control	Vision Loss	Parkinson's Disease	Ataxia	Downbeat Nystagmus
Number of chunks	16660	10165	26458	10374	14041

Our algorithm first processes short segments of a recording (chunks). A chunk is measured by its number of fixations (we use 10 fixations). This number of fixations is chosen to correspond to roughly 3 seconds of time. For each fixation, we generate 9 features:

1. Fixation X and Y position in pixels – the 2D location of the patient's gaze on the screen.
2. Fixation duration in milliseconds.
3. Fixation dispersion in pixels – X and Y axis length of an ellipse encompassing the patient's gaze locations associated with the fixation.
4. Number of saccades occurring directly after the fixation (before the subsequent fixation).
5. Total duration of saccades associated with the fixation in milliseconds.
6. X and Y position of the nearest target location to the fixation in pixels – this is our best

guess of where the person "should have" been looking to complete the given task.

We concatenate these features across time resulting in a $9 \times T$ feature vector where T is the number of fixations we use (e.g., 10 fixations). This representation is input into a convolutional neural network (CNN) which we train to predict disease group using the standard cross entropy loss. We refer to this CNN as our chunk-level classifier. During training, we augment our dataset by performing small, Gaussian perturbations to our chunk features. We use the Adam optimizer[20] and a cross-entropy loss to train our model. The CNN model is implemented in PyTorch.[21] It consists of 3 convolutional layers and 4 linear layers followed by a final linear classifier that outputs the logits used for classification. We use ReLU nonlinearities.

Note that a recurrent neural network (RNN) is an alternative model choice. However, we opted to use a CNN model for two key reasons: (1) we are interested in the chunk-level performance of our model, which precludes the use of any recurrence between chunks, and (2) we have enough data to train a CNN model that takes as input an entire chunk, and hence do not need an RNN model. For aggregating across chunks, an RNN is a viable candidate model that has seen prior use.[22] However, we opt for the simpler approach described below.

3.2.2. *Task-level classifier*

Although we trained the chunk-level classifier to identify disease, we will only use it to produce embeddings. In particular, we use the second to last layer output of the chunk-level classifier as a chunk-level embedding, $e_{\text{chunk}} \in \mathbb{R}^8$. For each recording, we compute this embedding for all possible chunks inside the given recording. Chunks are spaced with an offset of 1 fixation, resulting in a total of $N - T + 1$ chunks where N is the number of fixations in the recording and T is the length of the chunk. These embeddings are aggregated by computing the max, min, and average value for all embeddings for the recording and concatenating, resulting in a chunk-aggregate embedding $e_{\text{chunk-agg}} \in \mathbb{R}^{24}$.

In parallel, we generate 16 features at the task level. These features include total duration of the recording, total number of fixations, average distance from fixation to target points (e.g., the locations of the numbers to be read), and average fixation duration. We concatenate these features to the chunk-aggregate features to create our task-level embedding $e_{\text{task}} \in \mathbb{R}^{40}$. We train our task-level classifier, a random forest (RF), using cross validation for hyperparameter tuning to predict disease group from e_{task}.

4. Results

We first present results for our chunk-level classifier and discuss how we chose the chunk length. These results demonstrate that very short recordings are sufficient for disease detection. Subsequently, we discuss the task-level classifier and provide insights that can help future task design be more informative for disease detection and differentiation.

4.1. *Chunk-level classifier*

We first perform an analysis of the chunk length. In Figure 3a, we plot the ROC-AUC for control vs. disease prediction as a function of chunk length. Performance is paralleled across the

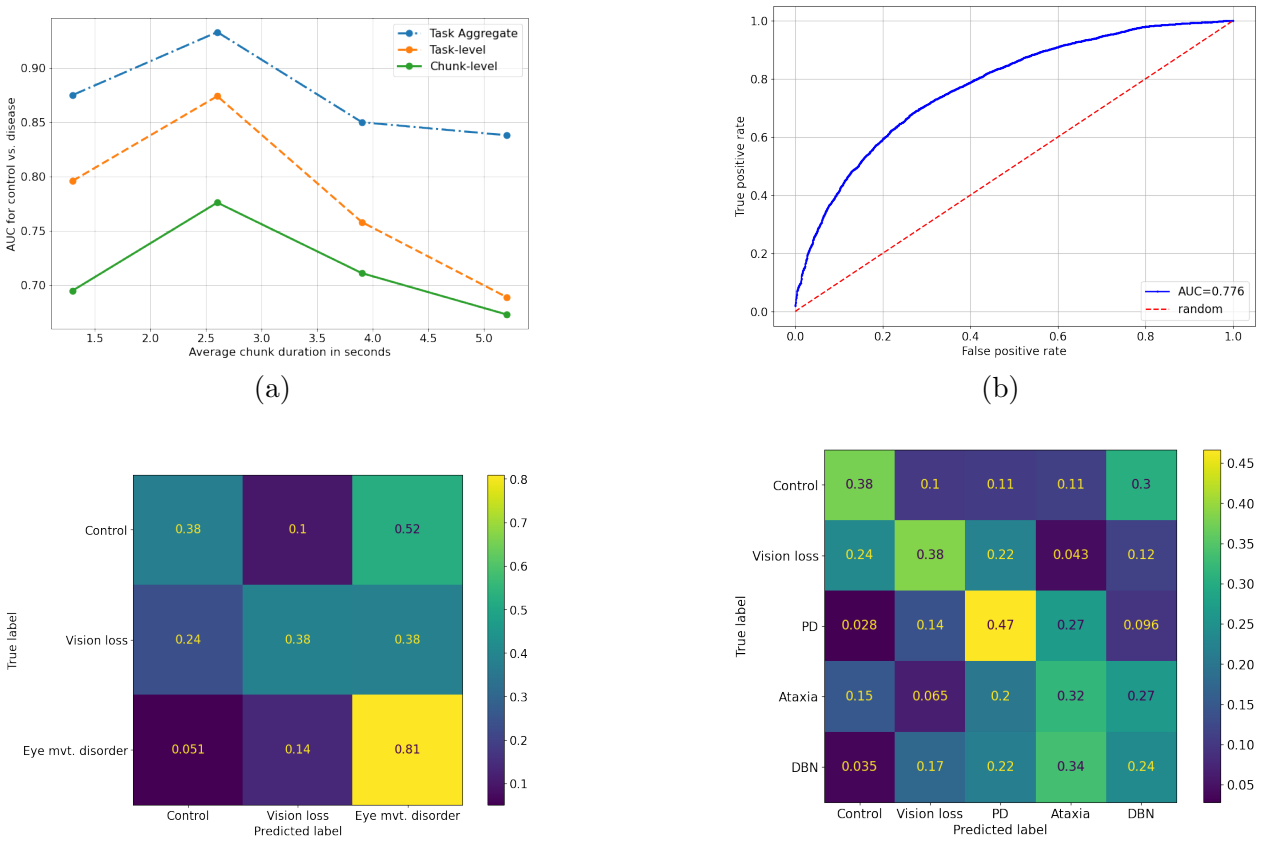


Fig. 3. Results for chunk-level classifier that predicts disease state from < 3 seconds of data. (a) ROC-AUC (control vs. disease) as a function of chunk length. We translate chunk length to the corresponding average chunk duration across our dataset, and note a peak at 2.6 seconds. (b) ROC-AUC plot for chunk-level classifier showing discrimination between healthy and diseased patients despite the limited input data. (c, d) Chunk-level classifier confusion matrix between (b) control state, vision loss, and eye movement disorders and (c) all disease states.

chunk-level, task-level, and task aggregate (majority vote across tasks) classifiers, suggesting performance is heavily reliant on the information extracted at the chunk level. Furthermore, this result suggests that the short recording segments are informative of disease state.

Performance is maximized at around 2.6 seconds average chunk duration. While we might expect increasing chunk length to only increase performance, there is an implicit tradeoff – as chunk length increases, model complexity increases while number of data points decreases. A shorter chunk length has higher performance due in part to overfitting resulting from our small dataset size. Based on these results, we use $T = 10$ fixations for each chunk.

Now we discuss the performance of our chunk-level classifier using the chunks of average duration 2.6 seconds. We show an ROC-AUC plot for classifying disease vs. not disease (i.e., aggregating all four disease states into one group) in Figure 3b, and attain an ROC-AUC value of 0.776.

In Figure 3c, we show a confusion matrix for the 3-way disease classification, and in Figure 3d, we show the confusion matrix for the full 5-way disease classification. While the correct classes are generally the most common prediction, there is a high degree of variance in predictions. *This is not unexpected:* the recordings are noisy due to both the eye tracker and patient variability, and signal is limited due to the short duration of the chunks.

4.2. *Task-level classifier*

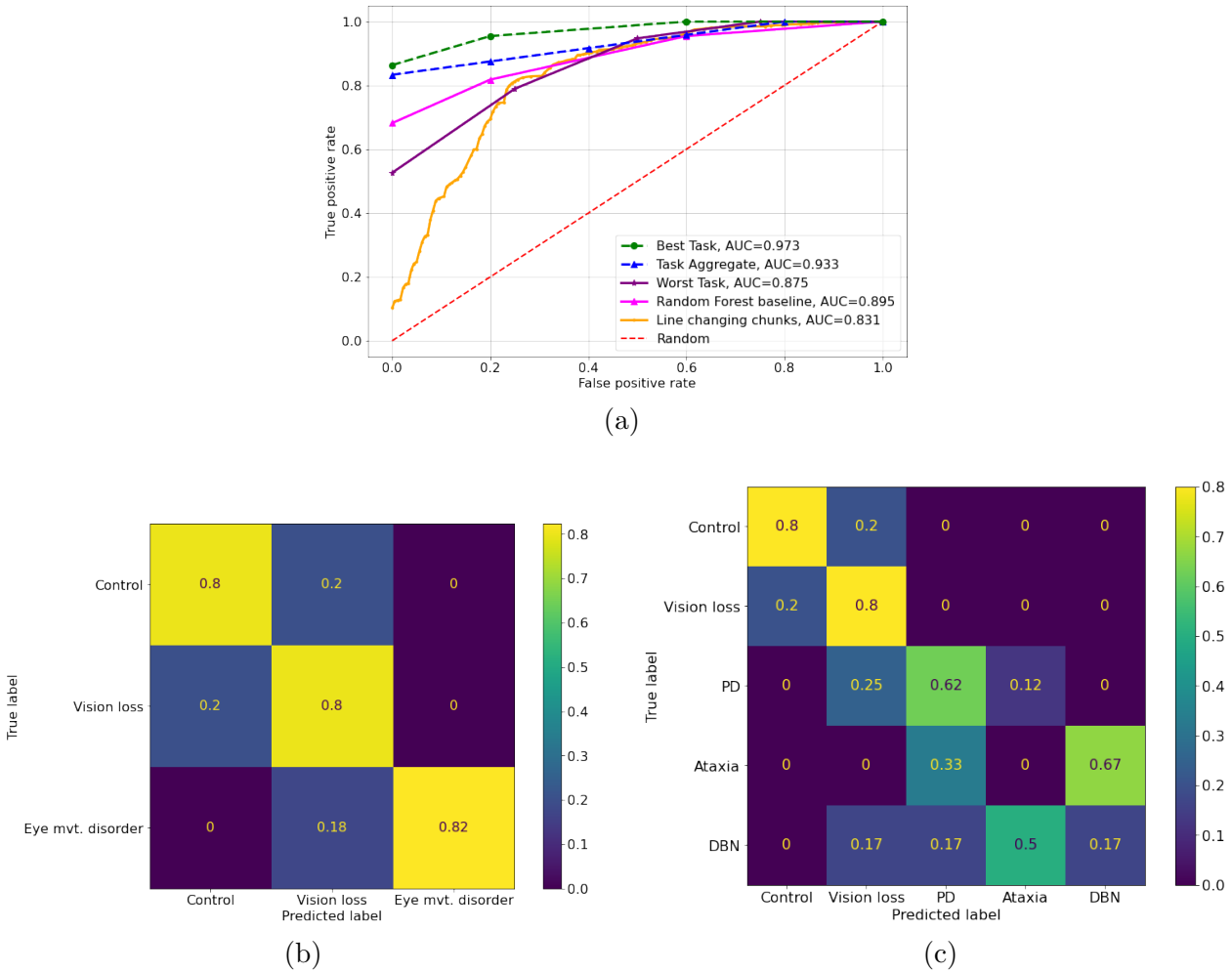
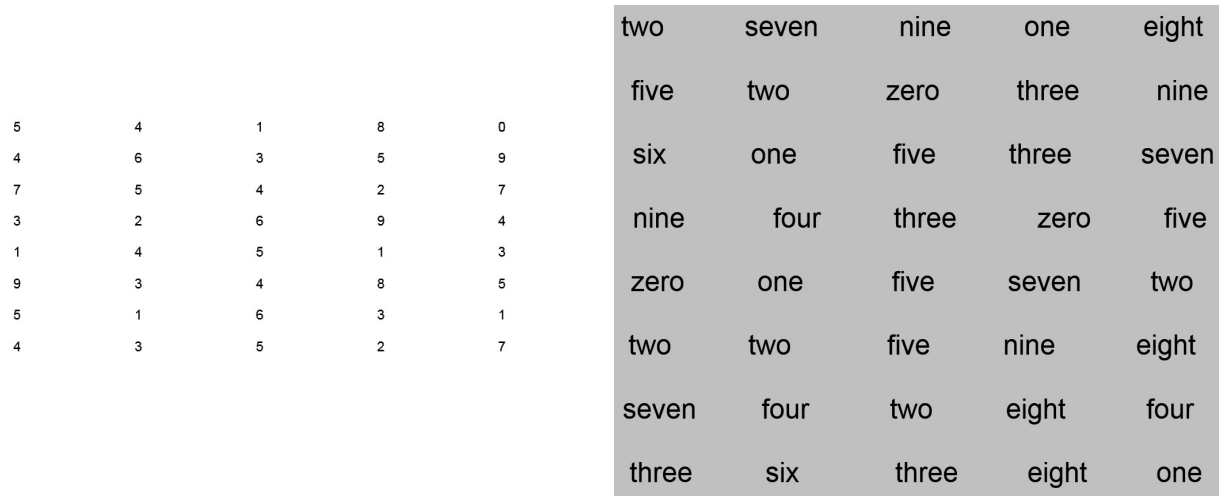


Fig. 4. Results for task-level disease classifier. (a) AUC plot comparing task-level classifier on best and worst tasks, aggregate by majority vote of task-level classifications, random forest trained on hand-crafted features on the best task, and the chunk-level classifier on line-changing chunks. (b, c) Task-level classifier confusion matrix for best task between (b) control state, vision loss, and eye movement disorders and (c) all disease states.

Here we present results for the task-level classifier. In Figure 4a, we compare ROC-AUC curves for the binary disease detection task for several different classifiers. The "Best Task" (see Figure 5a) and "Worst Task" (see Figure 5b) classifiers are the task-level classifier applied



(a) (b)

Fig. 5. (a) Best task (read right-to-left) and (b) worst task (read left-to-right), as measured by ROC-AUC value in Figure 4a. There is less spacing between lines in (a) than the example given in Figure 1a.

to only a single task across patients. The large performance gap suggests that certain tasks are indeed more informative than others. The "Task Aggregate" classifier is simply a majority vote for disease prediction across all tasks. Perhaps surprisingly, this aggregate prediction has lower performance than the best task, again suggesting the large performance variation between tasks. The "Random Forest baseline" is an RF model that used the hand-selected features without the CNN embeddings. It achieves 3-way accuracy of 46% on the best task and 49% when aggregated with majority vote across all tasks (not plotted here), which is substantially worse that our model (81% accuracy). This suggests the CNN embeddings contribute useful signal. The "Line changing chunks" model refers to the chunk-level classifier applied to the best task and only for the recording segments that capture how a patient moves from one line to the next while reading. The higher ROC-AUC value as compared to the ROC-AUC value in Figure 3b, which averages ROC-AUC across all chunks and all tasks, suggests that certain segments of the recording, like the line-changing segments, are more informative. We did not includes baselines from prior work as our setting is not directly comparable (for example, data in some prior work does not come from a reading task, and so direct adaptation to our setting is not possible).

In Figure 4b and Figure 4c, we show the confusion matrices for the 3-way and 5-way classification for the task-classifier applied only to the best task. The 3-way classifier achieves 81% accuracy against a random guessing baseline of 33%. Our 5-way classifier achieves 47% accuracy; notably, the classifier confuses Ataxia and DBN with one another. This is not unsurprising – as previously noted, PD, Ataxia, and DBN are similar due to the similar ways in which symptoms manifest in visuo-motor function.

Now comparing across all 12 tasks, we make a few observations. The tasks with higher visual crowding and tasks where we require reading right-to-left as opposed to left-to-right

both achieve higher ROC-AUC values (both increase difficulty), suggesting that a certain amount of difficulty is critical for task design. Additionally, the tasks requiring reading single digits as opposed to words had higher performance, indicating that the single-digit reading task may be more informative. Other tasks perform better at differentiating Ataxia and DBN, however, which may indicate that certain tasks are better at differentiating different diseases. Given our limited dataset size, it is difficult to make definitive conclusions.

5. Discussion

In this work, we demonstrated that (1) it is possible to use very short eye tracking recordings to accurately identify patients with abnormal eye movements due to a disease, and (2) with longer recordings, it is possible to differentiate between up to 5 different disease states with significantly higher accuracy than random chance. We identified paradigms important for task design that produce more informative data for disease detection and differentiation. Moreover, our end-to-end ML model provides a framework to leverage eye-movements at different temporal resolutions to classify visuo-motor diseases.

Future work includes gathering additional data for ensuring model robustness across various confounders including age and gender. Additionally, we believe our approach can be extended to characterize disease progression by utilizing the chunk and stimuli-level embeddings for unsupervised analysis. Our work is highly relevant to the medical community. Recent advances in eye-tracking on smartphone cameras have been shown to approach clinical grade eye tracker performance.[23] A solution like ours that can identify disease using only short recordings of eye movement data could be combined with the new eye-tracking approaches to provide automatic detection for various diseases across a global population.

References

1. P. S. Holzman, L. R. Proctor and D. W. Hughes, Eye-tracking patterns in schizophrenia, *Science* **181**, 179 (1973).
2. J. T. Hutton, J. Nagel and R. B. Loewenson, Eye tracking dysfunction in alzheimer-type dementia, *Neurology* **34**, 99 (1984).
3. E. Pretegiani and L. M. Optican, Eye movements in parkinson's disease and inherited parkinsonian syndromes, *Frontiers in neurology* **8**, p. 592 (2017).
4. N. Jehangir, C. Y. Yu, J. Song, M. A. Shariati, S. Binder, J. Beyer, V. Santini, K. Poston and Y. J. Liao, Slower saccadic reading in parkinson's disease, *Plos one* **13**, p. e0191005 (2018).
5. M. Hunfalvay, C.-M. Roberts, N. Murray, A. Tyagi, H. Kelly and T. Bolte, Horizontal and vertical self-paced saccades as a diagnostic marker of traumatic brain injury, *Concussion* **4**, p. CNC60 (2019).
6. A. King, The proposed king-devick test and its relation to the pierce saccade test and reading levels, *Available from the Carl Shepherd Memorial Library, Illinois College of Optometry, Chicago, IL* (1976).
7. G. T. Liu, N. J. Volpe and S. L. Galetta, The neuro-ophthalmic examination, in *Liu, Volpe, and Galetta's Neuro-Ophthalmology*, (Elsevier, 2019) pp. 7–36.
8. T. Ilginis, J. Clarke and P. J. Patel, Ophthalmic imaging., *British medical bulletin* **111** (2014).
9. C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel *et al.*, Movement disorder society-sponsored

revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results, *Movement disorders: official journal of the Movement Disorder Society* **23**, 2129 (2008).
10. L. G. C. Bautista and P. C. Naval, Gazemae: general representations of eye movements using a micro-macro autoencoder, in *2020 25th International Conference on Pattern Recognition (ICPR)*, (IEEE, 2021).
11. J. Kacur, J. Polec, E. Smolejova and A. Heretik, An analysis of eye-tracking features and modelling methods for free-viewed standard stimulus: Application for schizophrenia detection, *IEEE Journal of Biomedical and Health Informatics* **24**, 3055 (2020).
12. K. A. Dalrymple, M. Jiang, Q. Zhao and J. T. Elison, Machine learning accurately classifies age of toddlers based on eye tracking, *Scientific reports* **9**, 1 (2019).
13. R. C. Shaffer, E. V. Pedapati, F. Shic, K. Gaietto, K. Bowers, L. K. Wink and C. A. Erickson, Brief report: diminished gaze preference for dynamic social interaction scenes in youth with autism spectrum disorders, *Journal of autism and developmental disorders* **47**, 506 (2017).
14. M. S. Daley, D. Gever, H. F. Posada-Quintero, Y. Kong, K. Chon and J. B. Bolkhovsky, Machine learning models for the classification of sleep deprivation induced performance impairment during a psychomotor vigilance task using indices of eye and face tracking, *Frontiers in Artificial Intelligence* **3**, p. 17 (2020).
15. T. Kootstra, J. Teuwen, J. Goudsmit, T. Nijboer, M. Dodd and S. Van der Stigchel, Machine learning-based classification of viewing behavior using a wide range of statistical oculomotor features, *Journal of vision* **20**, 1 (2020).
16. Y. Yamada and M. Kobayashi, Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults, *Artificial intelligence in medicine* **91**, 39 (2018).
17. A. L. M. Pambakian, D. Wooding, N. Patel, A. Morland, C. Kennard and S. Mannan, Scanning the visual world: a study of patients with homonymous hemianopia, *Journal of Neurology, Neurosurgery & Psychiatry* **69**, 751 (2000).
18. C. Mariotti, R. Fancellu and S. Di Donato, An overview of the patient with ataxia, *Journal of neurology* **252**, 511 (2005).
19. R. D. Yee, Downbeat nystagmus: characteristics and localization of lesions., *Transactions of the American Ophthalmological Society* **87**, p. 984 (1989).
20. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, eds. Y. Bengio and Y. LeCun (ICLR, 2015).
21. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (Curran Associates, Inc., 2019) pp. 8024–8035.
22. Y. Mao, Y. He, L. Liu and X. Chen, Disease classification based on eye movement features with decision tree and random forest, *Frontiers in Neuroscience* **14**, p. 798 (2020).
23. N. Valliappan, N. Dai, E. Steinberg, J. He, K. Rogers, V. Ramachandran, P. Xu, M. Shojaeizadeh, L. Guo, K. Kohlhoff *et al.*, Accelerating eye movement research via accurate and affordable smartphone eye tracking, *Nature communications* **11**, 1 (2020).